João Vítor Possamai de Menezes

Uma análise audiovisual da produção de tons lexicais

Belo Horizonte

2020

João Vítor Possamai de Menezes

Uma análise audiovisual da produção de tons lexicais

Dissertação apresentada ao Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Minas Gerais como requisito parcial para obtenção do título de Mestre em Engenharia Elétrica.

Orientador: Prof. Dr. Adriano Vilela Barbosa

Coorientador: Profa. Dra. Maria Mendes Cantoni

Belo Horizonte

2020

Menezes, João Vítor Possamai de.
M543u Uma análise audiovisual da produção de tons lexicais [recurso eletrônico] / João Vítor Possamai de Menezes. – 2020.
1 recurso online (112 f. : il., color.) : pdf.

Orientador: Adriano Vilela Barbosa.
Coorientadora: Maria Mendes Cantoni.

Dissertação (mestrado) Universidade Federal de Minas Gerais, Escola de Engenharia.

Bibliografia: f. 104-112.
Exigências do sistema: Adobe Acrobat Reader.

1. Engenharia elétrica - Teses. 2. Fala - Teses. Lexicologia I. Barbosa, Adriano Vilela. II. Cantoni, Maria Mendes. III. Universidade Federal de Minas Gerais. Escola de Engenharia. IV. Título.

CDU: 621.3(043)

Ficha catalográfica: Biblioteca Prof. Mário Werneck, Escola de Engenharia da UFMG

"Uma Análise Audiovisual da Produção de Tons Lexicais"

João Vítor Possamai de Menezes

Dissertação de Mestrado submetida à Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em Engenharia Elétrica da Escola de Engenharia da Universidade Federal de Minas Gerais, como requisito para obtenção do grau de Mestre em Engenharia Elétrica.

Aprovada em 31 de julho de 2020.

Por:

Prof. Dr. Adriano Vilela Barbosa
DELT (UFMG) - Orientador

Prof. Dra. Maria Mendes Cantoni
FALE (UFMG) - Co-orientadora

Prof. Dr. Hani Camille Yehia
DELT (UFMG)

Prof. Dr. Frederico Gualberto Ferreira Coelho
DELT (UFMG)

Prof. Dr. Adriano Chaves Lisboa
Gaia Soluções em Engenharia

Este trabalho é dedicado às pessoas que dedicam seu tempo e sua saúde mental e física à defesa da educação pública, gratuita e de qualidade no Brasil.

Agradecimentos

Começo agradecendo às pessoas que desde sempre estiveram a meu lado: minha mãe Karina e meu pai José Vitor. Criaram-me com amor e respeito, e sempre pelo exemplo. Aos dois a minha admiração e o meu amor eternos. A vida avança e nos proporciona novas companhias. A primeira que recebi foi a do meu irmão, Leonardo, por quem sou muito grato. A ele minha admiração e meu apoio incondicional. A vida e as pessoas não surgem simplesmente: são cultivadas. Agradeço também às pessoas que cultivaram minha família: minhas avós Isolde e Marly e meu avô Oswaldo. Sou muito grato pelos sentimentos e momentos que compartilhamos na nossa convivência.

Outra companhia maravilhosa que a vida me proporcionou foi a Karoline, minha grande companheira. A você e por você agradeço todos os dias. Você me motiva e me fascina. Me considero privilegiado por ter você na minha vida e também por fazer parte da sua. Agradeço também às pessoas que fizeram com que eu me sentisse, em Belo Horizonte e em Divinópolis, quase tão em casa quanto em Curitiba: Donizeti, Kamila e Marlene. Sou muito grato pelo amor e carinho de vocês, e também pelo o que aprendi com vocês.

Mais companhias que a vida me proporcionou e que, pela longevidade das nossas relações, já são parte de mim: Enrico, Franco, Jacqueline, Rafael, Sidnei, Thiago e Vitor. Admiro muito vocês e quero vocês na minha vida sempre.

Ao meu orientador Adriano e à minha coorientadora Maria os meus sinceros agradecimentos. Vocês me orientaram não só dedicando seus intelectos e seu tempo, mas também pelo exemplo de profissionais e pessoas que são. Fui privilegiado por poder desenvolver este trabalho com vocês. Ao professor Hani, meus agradecimentos pelas orientações pessoais e profissionais e por acreditar e impulsionar a pesquisa multidisciplinar na engenharia. Sou grato por me sentir parte do CEFALA (Centro de Estudos da Fala, Acústica, Linguagem e músicA). Ao professor Denis Burnham, meus agradecimentos pela confiança depositada em nós para trabalharmos com dados tão valiosos quanto os que você e sua equipe coletaram. Agradeço também às pessoas cujas trajetórias de encontraram temporal e espacialmente com a minha, pois se tornaram companheiros e companheiras: Adrielle, Carla, Felipe, Gabriela, Leandro, Melchior e Roger.

Por fim, agradeço à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo apoio financeiro durante o período de setembro de 2018 até julho de 2020 sob o processo 88882.381124/2019 − 1. Você é essencial para o Brasil.

Resumo

Sabe-se que a fala se manifesta não só de forma acústica, mas também visual, por meio de movimentos faciais e gestos corporais, além de possuir correlatos fisiológicos como o movimento do trato vocal e a atividade neural. Este trabalho apresenta uma análise audiovisual da produção de tons lexicais, que são variações de graves e agudos que mudam o significado das palavras em línguas tonais. Tons lexicais são tradicionalmente estudados em termos de parâmetros acústicos, como a frequência fundamental (F0) do sinal de fala. Este trabalho, no entanto, adota uma abordagem integrada, investigando a contribuição, de forma isolada e conjunta, das componentes acústica e visual da fala para a diferenciação dos tons lexicais em três línguas tonais (cantonês, mandarim e tailandês). A abordagem adotada é tentar classificar os tons de cada língua a partir de cada componente tomada isoladamente e comparar seus desempenhos. Foram coletados dados em experimentos audiovisuais de produção de fala com sete falantes das três línguas. A componente visual da fala foi obtida por meio do rastreamento 3D de marcadores fixados à face e à cabeça das participantes, e a componente acústica foi obtida, de forma simultânea, por um microfone. Após o experimento, as posições dos marcadores foram submetidas a um procedimento de compensação do movimento da cabeça com o intuito de decompô-las em suas duas componentes: uma devida ao movimento da face e outra devida ao movimento de corpo rígido da cabeça. O sinal acústico teve sua F0 estimada por meio do método de autocorrelação. Neste ponto, a componente visual é representada por três tipos de sinais: Movimento Total (posições dos marcadores), Face e Cabeça (resultantes da decomposição); e a componente acústica é representada pelas curvas de F0. Todos os tipos de sinais foram parametrizados por meio de regressão polinomial, sendo representados por coeficientes que aproximam sua trajetória original. Os sinais parametrizados foram então utilizados para treinar classificadores lineares e não-lineares, com os tons de cada língua usados como rótulos das classes. A capacidade de cada tipo de sinal de classificar os diferentes tons lexicais foi medida por meio da acurácia de cada classificador, obtida com validação cruzada em K partes ($K = 5$). Os sinais visuais foram capazes de classificar tons lexicais, nas três línguas, com acurácia acima da aleatória. As maiores acurácias foram obtidas pelos sinais de F0. Entre os sinais visuais, as maiores acurácias foram obtidas, em ordem decrescente, pelos sinais Movimento Total e Face. Além disso, alguns tons lexicais de uma mesma língua foram classificados com acurácias acima da média, sugerindo que alguns tons são mais fáceis de serem classificados do que outros. Os resultados obtidos estão de acordo com a literatura e sugerem que tons lexicais podem ser preditos não só por F0, mas também, em menor grau, pelos movimentos da face e da cabeça.

Palavras-chave: Fala multimodal, Línguas tonais, Tom lexical, Classificação estatística

Abstract

It is known that speech manifests itself not only acoustically, but also visually, through facial movements and body gestures, in addition to having physiological correlates such as movement of the vocal tract and neural activity. This work presents an audiovisual analysis of the production of lexical tones, which are pitch variations that change the meaning of words in tone languages. Lexical tones are traditionally studied in terms of acoustic parameters, such as the fundamental frequency (F0) of the speech signal. This work, however, adopts an integrated approach, investigating the contribution, in isolation and jointly, of the acoustic and visual components of speech to the differentiation of lexical tones in three tone languages (Cantonese, Mandarin and Thai). The approach adopted consists in classifying the tones of each language from each component taken in isolation and to compare their performances. Data was collected in audiovisual speech production experiments with seven speakers of the three languages. The visual component of speech was obtained through 3D tracking of markers fixed to the participants' faces and heads, and the acoustic component was obtained simultaneously by a microphone. After the experiment, the positions of the markers were subjected to a head movement compensation procedure in order to separate them into their two components: one due to the movement of the face and the other due to the movement of the rigid body of the head. The acoustic signal had its F0 estimated through the autocorrelation method. At this point, the visual component is represented by three types of signals: Total movement (marker positions), Face and Head (resulting from the decomposition); and the acoustic component is represented by the F0 curves. All types of signals were parameterized using polynomial regression, being represented by coefficients that approximate their original trajectory. The parameterized signals were then used to train linear and non-linear classifiers, with the tones of each language used as class labels. The ability of each type of signal to classify the different lexical tones was measured using the accuracy of each classifier, obtained with cross-validation in K parts ($K = 5$). Visual signals were able to classify lexical tones in the three languages, with accuracy above chance. The highest accuracy was obtained by the F0 signals. Among the visual signals, the highest accuracy was obtained, in decreasing order, by the signals Total Movement and Face. In addition, some lexical tones of the same language were classified with above-average accuracy, suggesting that some tones are easier to classify than others. The results obtained are in accordance with the literature and suggest that lexical tones can be predicted not only by F0 , but also, to a lesser extent, by the movements of the face and head.

Keywords: Multimodal speech, Tone languages, Lexical tone, Statistic classification.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

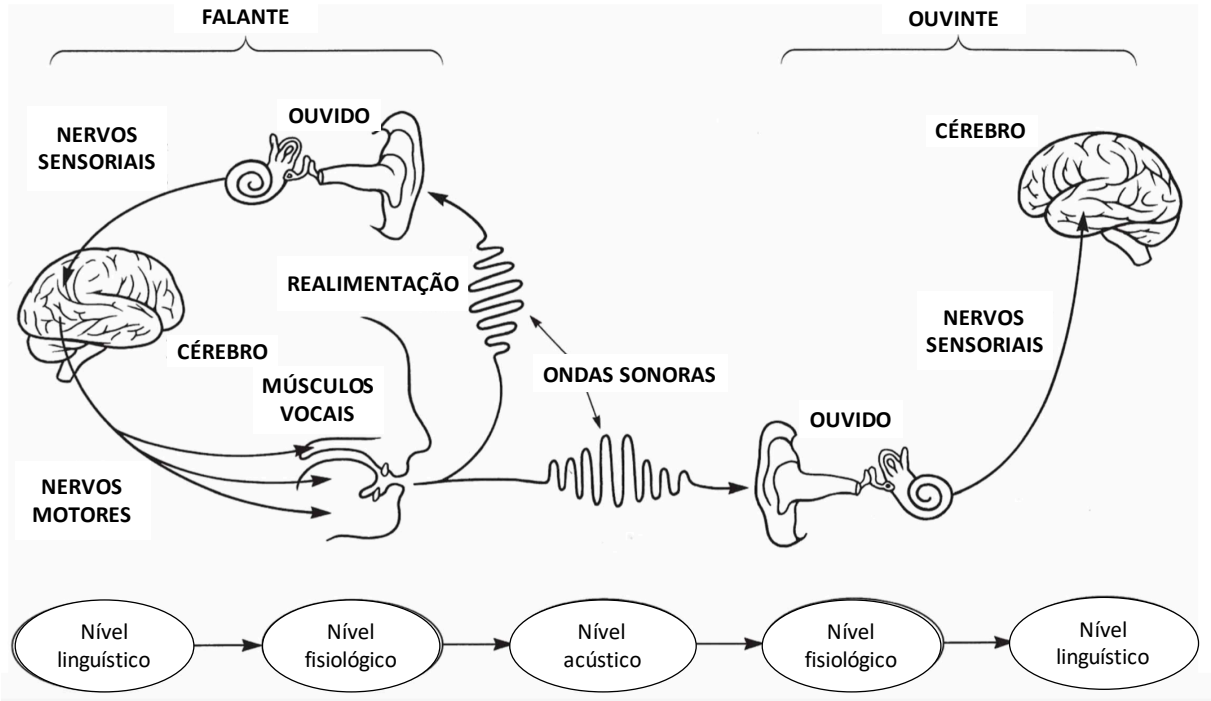
A fala é um fenômeno complexo que depende de diferentes sistemas do corpo humano para sua realização. Uma forma simples e didática de caracterizá-la é através do modelo da Cadeia da Fala, composta pelos níveis linguístico, fisiológico e acústico (DENES; PINSON, 2015). O fenômeno da fala se inicia no nível linguístico, em que o falante define uma mensagem a ser transmitida, selecionando as palavras e ordenando-as de modo adequado. Já no nível fisiológico, seu cérebro envia impulsos nervosos aos músculos que ativam o aparelho fonador, produzindo ondas sonoras. O nível acústico consiste na propagação dessa onda sonora desde as pregas vocais, passando pelo trato vocal, sendo irradiada pelas cavidades oral e nasal, e continuando no espaço até o aparelho auditivo do ouvinte. Agora, com a participação do ouvinte, tem-se o processo inverso. No nível fisiológico, os mecanismos de audição atuam transformando as ondas sonoras em impulsos nervosos e, no nível linguístico, a mensagem é finalmente reconhecida. A essa cadeia de eventos, adiciona-se uma retroalimentação das ondas sonoras produzidas pelo falante por meio da própria audição. Isso implica que a percepção do falante sobre o que ele mesmo fala é um fator que influencia o processo de produção da fala. A Cadeia da Fala aqui descrita é ilustrada na Figura 1.

A Cadeia da Fala, contudo, é um modelo que contempla apenas sua componente acústica. Se a fala for vista sob o aspecto de uma interação falante-ouvinte condicionada à transmissão de significado, então ela não deve ser caracterizada apenas por fenômenos acústicos, mas por quaisquer fenômenos capazes de transmitir significado, tais como gestos articulatórios (BROWMAN; GOLDSTEIN, 1986; SALTZMAN; MUNHALL, 1989), corporais (MCNEILL, 1981) e faciais (SUMBY; POLLACK, 1954; MCGURK; MACDONALD, 1976), que se manifestam visualmente.

Como a fala se manifesta por meio de diferentes modalidades, como a acústica e a visual, pode-se caracterizá-la como um fenômeno multimodal. A partir da década de 50, trabalhos começaram a evidenciar a relevância da modalidade visual da fala por meio dos gestos faciais. Entre os resultados mais importantes, pode-se citar que a inteligibilidade da fala aumenta significativamente quando o ouvinte possui pistas visuais em adição às acústicas (SUMBY; POLLACK, 1954) e que disparidades entre estímulos visuais e auditivos alteram o significado percebido pelo ouvinte (MCGURK; MACDONALD, 1976).

A década de 80 foi o momento do surgimento de várias teorias que buscam associar outros gestos à produção e à percepção da fala, tendo destaque entre eles a Fonologia Articulatória (FA) (BROWMAN; GOLDSTEIN, 1986; SALTZMAN; MUNHALL, 1989). Procurando relacionar as dimensões físicas e fonológicas da fala, a FA propõe uma

Figura 1 – A Cadeia da Fala.



Fonte: Adaptado de (DENES; PINSON, 2015)

representação da fala baseada na sobreposição espacial e temporal de padrões de gestos articulatórios, como o movimento dos lábios, da língua e do maxilar, por exemplo. Para a FA, os gestos são padrões abstratos que podem ser efetivamente atingidos ou não durante a fala. Outros estudos também se ocuparam do papel dos gestos corporais, como movimentos dos braços e das mãos, na fala (MCNEILL, 1981). Assim como as ondas sonoras, esses gestos têm a capacidade de transmitir significado, sendo as línguas de sinais um exemplo prático de transmissão de significado por meio de movimentos corporais.

Há ainda outras partes da componente visual da fala que se relacionam diretamente com a componente acústica, como o movimento do trato vocal, que é fisiologicamente necessário para a produção da fala. Resultados importantes demonstraram a relação do movimento do trato vocal com a componente acústica e também com outras componentes visuais da fala (MERMELSTEIN, 1967; YEHIA; RUBIN; VATIKIOTIS-BATESON, 1998). O formato do trato vocal pode ser estimado com base nas frequências formantes do sinal acústico, como realizado por (MERMELSTEIN, 1967). Além disso, por meio de uma série de estimadores lineares, foi possível estimar diferentes dimensões da fala, uma a partir da outra, como realizado por (YEHIA; RUBIN; VATIKIOTIS-BATESON, 1998). Nesse trabalho, o movimento do trato vocal foi obtido a partir do movimento orofacial (movimentos da face e da boca) e vice-versa por meio de estimadores lineares. Além disso, parâmetros acústicos da fala foram obtidos a partir do movimento do trato vocal

também por meio de estimadores lineares. Os estudos citados acima demonstraram relações quantitativas entre as diferentes modalidades da fala, contribuindo para o entendimento dos mecanismos de produção da fala e ampliando as possibilidades de estudos na área.

Para a obtenção de dados de movimento durante a fala são necessários equipamentos de medição específicos. Entre eles, destacam-se pelo grande número de aplicações o OPTOTRAK (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020a) e o articulógrafo eletromagnético (EMA, do inglês *Electromagnetic Articulograph*) (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020c; KARSTENS MEDIZINELEKTRONIK GMBH, 2020).

O OPTOTRAK é um equipamento de rastreamento de marcadores emissores de luz infravermelha que são, geralmente, posicionados na face do falante de modo a capturar os movimentos da face e da cabeça (VATIKIOTIS-BATESON; OSTRY, 1995). O EMA obtém um sinal que representa o movimento do trato vocal a partir do rastreamento magnético de pequenos sensores (bobinas) colocados em pontos como a língua, os lábios e a mandíbula, por exemplo.

Transmissores de rádio-frequência criam um campo eletromagnético no espaço onde os sensores se encontram, e é induzida corrente elétrica nos sensores quando eles se movem na presença do campo eletromagnético. A posição dos sensores é então determinada a partir das correntes que foram induzidas, comparando-as com correntes de referência (PERKELL et al., 1992; TIEDE et al., 2012). Uma outra técnica também utilizada é a eletromiografia (EMG), que monitora a atividade elétrica das células musculares, capturando os padrões de ativação de músculos durante a realização de movimento (ZAHNER et al., 2014; DIENER et al., 2016).

As técnicas apresentadas acima impulsionaram o desenvolvimento da pesquisa em fala multimodal por possibilitarem o rastreamento de componentes visuais da fala com boa precisão espacial e temporal. Enquanto o modelo mais atual de OPTOTRAK possui precisão espacial de até $0,1$mm e frequência de amostragem dos quadros igual a 4600Hz/(número de marcadores $+ 1,3$) (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020a), os modelos de EMA possuem resolução espacial de $0,9$mm e frequência máxima de amostragem de 400Hz (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020c). Essas técnicas possuem, por outro lado desvantagens que dificultam sua aplicação. Medições com EMA e EMG são invasivas, pois há posicionamento de sensores dentro da cavidade oral (EMA) e abaixo da pele (EMG), necessitando de profissionais especializados para esse procedimento. Além disso, OPTOTRAK, EMA e EMG necessitam de estrutura auxiliar em condições de laboratório, como a instalação de vários condutores ligando os sensores ao equipamento, o que dificulta o transporte da configuração do experimento. Toda a estrutura necessária para a aplicação dessas técnicas também limita a movimentação do falante, afastando as condições de produção da fala da naturalidade, que é a condição

ideal. Outro fator negativo da aplicação dessas técnicas é seu alto custo, pois dependem de equipamentos específicos e de alta precisão.

Uma alternativa a essas técnicas é o uso de técnicas não-invasivas baseadas em vídeo, como o fluxo óptico, que quantifica o movimento capturado em vídeo. Existem vários tipos de algoritmos que implementam o fluxo óptico, mas o objetivo básico é gerar um campo 2D de movimento a partir de padrões espaço-temporais de intensidade das imagens (BARRON; FLEET; BEAUCHEMIN, 1994). A liberdade de poder capturar dados apenas com uma câmera, já que o fluxo óptico consiste em pós-processamento de vídeo, é uma grande vantagem em relação ao OPTOTRAK, ao EMA e ao EMG, que dependem de condições de laboratório para sua utilização. O fluxo óptico, por outro lado, não oferece tanta precisão quanto os outros métodos (BARRON; FLEET; BEAUCHEMIN, 1994), sendo a escolha pela técnica de medição em experimentos de fala multimodal muito influenciada pelo compromisso entre precisão e facilidade/custo.

O uso integrado de mais de uma das técnicas de aquisição de sinais apresentadas acima é comum em trabalhos da área de fala multimodal, que buscam avaliar a relação entre diferentes modalidades da fala. Entre vários estudos que aplicam duas ou mais dessas técnicas em conjunto, pode-se citar o trabalho de (YEHIA; RUBIN; VATIKIOTIS-BATESON, 1998), em que EMA e OPTOTRAK foram empregados para capturar os movimentos do trato vocal e da face, respectivamente; o trabalho de (VATIKIOTIS-BATESON; MUNHALL et al., 1996), em que as técnicas de EMG e de fluxo óptico foram combinadas para medir o movimento facial; e também, mais recentemente, o trabalho de (DANNER; BARBOSA; GOLDSTEIN, 2018), em que EMA e fluxo óptico foram combinados para a medição de movimentos corporais relacionados à fala.

Com a diversidade de técnicas de aquisição de sinais, surgem também modos de análise conjunta de sinais de diferentes modalidades, como a acústica e a visual, por exemplo. Um exemplo é a Análise por Mapa de Correlação (CMA, do inglês *Correlation Map Analysis*) (VILELA BARBOSA et al., 2012), que com base na correlação esperada entre os sinais da fala de modalidades diferentes, produz um gráfico 2D com a correlação instantânea entre eles em função do tempo e do atraso temporal, sendo uma útil ferramenta na análise da correlação entre dois sinais.

O campo de estudo que investiga a fala por meio de mais de uma modalidade é chamado de fala multimodal, cujos objetos de estudo podem ser a modalidade acústica da fala, o movimento orofacial, o movimento do trato vocal, o movimento corporal e a atividade neural. Devido à variedade dos sinais de entrada, a pesquisa em fala multimodal foi e ainda é impulsionada pelo desenvolvimento de técnicas de medição e análise que permitam a integração entre sinais de diferentes naturezas.

1.1 Motivação

Há diferentes aspectos linguísticos que recebem a atenção da pesquisa em fala multimodal, como a fala individual (DANNER; BARBOSA; GOLDSTEIN, 2018), a conversação entre falantes (LATIF et al., 2014) e a aquisição (SINGH; FU, 2016) e percepção (LIANG; HEUVEN, 2007; YUAN, 2011) de segunda língua, além de fenômenos linguísticos característicos de línguas específicas. Enquanto todas as línguas fazem uso de consoantes e vogais para compor o significado das palavras, existem línguas específicas chamadas de tonais que são caracterizadas pelo uso de **tons lexicais**, que são, de forma simples, variações de graves e agudos durante a pronúncia das palavras que alteram seu significado. Exemplos de línguas tonais são o cantonês, o mandarim e o yorubá (falado na África) (YIP, M., 2002). Por exemplo, a sequência de consoantes e vogais [yau] em cantonês pode ter 6 significados diferentes, cada um associado a um dos 6 tons lexicais que existem na língua. Por outro lado, em línguas não tonais, como o espanhol, o inglês e o português, essas variações de graves e agudos não determinam o significado das palavras, mas ainda podem exercer outras funções. Uma dessas funções é a entoação, que pode sinalizar, por exemplo, se a frase é uma pergunta, caso ela termine mais aguda, ou uma afirmativa, caso ela termine mais grave (MADDIESON, 2013). Uma caracterização mais detalhada de tons e línguas tonais é feita no Capítulo 3.

De particular importância para este trabalho são as línguas tonais, que são cerca de 70% de todas as línguas faladas na Terra, com centenas de milhões de falantes nativos (YIP, M., 2002). Além disso, a língua com o maior número de falantes nativos no mundo, o mandarim, é uma língua tonal (YIP, M., 2002). A contribuição que a informação visual (movimentos da face, gestos corporais etc.) traz para a inteligibilidade da fala, especialmente em situações em que há ruído, não foi observada apenas para o caso de consoantes e vogais, mas também para tons lexicais em línguas tonais como o cantonês e o mandarim (BURNHAM; CIOCCA; STOKES, 2001; BURNHAM; LAU et al., 2001; MIXDORFF; HU; BURNHAM, 2005; CHEN; MASSARO, 2008; SMITH; BURNHAM, 2012; GARG et al., 2019). Isso faz das línguas tonais um campo fértil para a pesquisa em fala multimodal, dado o grande número de falantes no mundo e o crescente número de estudos publicados na área. Além disso, as evidências de que o tom lexical é um fenômeno multimodal motivam pesquisas básicas e aplicadas que podem se relacionar a vários problemas, como por exemplo: 1) a percepção de tons lexicais por pessoas com deficiências auditivas, 2) a aquisição de línguas tonais por crianças com deficiências auditivas, 3) o desenvolvimento de aplicações de reconhecimento de fala audiovisual mais precisas e de síntese de fala mais realistas em línguas tonais, 4) o desenvolvimento de ferramentas de aprendizados de língua estrangeira mais específicos para línguas tonais.

Em estudos que realizaram experimentos de percepção com tarefas de diferenciação

dos tons lexicais do cantonês a partir de estímulos apenas visuais (VO)[1] compostos por movimentos dos lábios e da face, obteve-se desempenho acima do aleatório tanto por nativos (BURNHAM; CIOCCA; STOKES, 2001) quanto por não nativos (com uma língua nativa tonal, o tailandês, e uma língua nativa não-tonal, o inglês) (BURNHAM; LAU et al., 2001). Houve também resultados significativos com o mandarim, sugerindo ainda que há uma integração entre as informações acústicas e visuais em situações AV que auxilia na percepção dos tons lexicais na presença de ruído (MIXDORFF; HU; BURNHAM, 2005; SMITH; BURNHAM, 2012) e que há relação entre a produção de tons lexicais e os movimentos visíveis da cabeça, do pescoço e da boca do falante (CHEN; MASSARO, 2008). Mais recentemente, com o uso de técnicas de visão computacional na análise de dados de produção de tons lexicais, foram novamente sugeridas relações entre movimentos faciais, especificamente das sobrancelhas e dos lábios, com a produção de tons lexicais em mandarim (GARG et al., 2019).

Apesar disso, não há ainda consenso em relação à universalidade da multimodalidade da percepção de tons lexicais. Um motivo para isso é que os principais estudos no tema foram realizados com poucos falantes nativos produzindo estímulos (um falante em (BURNHAM; CIOCCA; STOKES, 2001), em (BURNHAM; LAU et al., 2001) e em (MIXDORFF; HU; BURNHAM, 2005); dois falantes em (SMITH; BURNHAM, 2012); quatro falantes em (CHEN; MASSARO, 2008); e vinte falantes em (GARG et al., 2019), que é a exceção nesta lista, com um número maior de falantes). Além disso, também há estudos cujos resultados sugerem que a percepção de tons lexicais não é sempre multimodal (HAN et al., 2019). Nesse estudo, por meio da comparação entre a integração audiovisual, numa situação AV, na percepção de tons lexicais em mandarim por falantes nativos e não-nativos, foram obtidos resultados sugerindo que falantes nativos utilizam apenas a componente acústica na percepção, enquanto que a componente visual é utilizada apenas por falantes não-nativos e em alguns tons específicos. Além disso, os estímulos para este estudo foram gravados por apenas uma falante nativa de mandarim.

Sabe-se, então, que a percepção de tons lexicais é baseada principalmente na informação acústica (HAN et al., 2019), e que os tons lexicais são geralmente qualificados em função de sua frequência fundamental (YIP, M., 2002), uma propriedade acústica. Sabe-se também que a informação visual contribui nessa percepção, com a maioria dos estudos que a investigam sendo de natureza qualitativa. Existem estudos que tentaram investigar essa contribuição de forma quantitativa, associando, com o auxílio de técnicas de visão computacional, movimentos específicos da face e da cabeça a tons lexicais específicos (GARG et al., 2019). No entanto, ainda há outros métodos quantitativos que podem

[1] Neste trabalho são citados resultados de experimentos de percepção que utilizaram, basicamente, três condições experimentais diferentes: 1) aquelas onde apenas estímulos acústicos são usados (AO - *Audio Only*), 2) aquelas onde apenas estímulos visuais são usados (VO - *Visual Only*) e 3) aquelas onde estímulos acústicos e visuais são usados (AV - *AudioVisual*).

ser aplicados a esse problema, verificando, por exemplo, a capacidade da informação visual de discriminar os tons lexicais uns dos outros. Além do pequeno número de estudos quantitativos, outras lacunas existentes na investigação desse problema são a falta de resultados que, por meio de diversos métodos, sugiram movimentos específicos da face e da cabeça sistematicamente relacionados a tons lexicais específicos e que tenham sido realizados com dados de produção de falantes nativos de diversas línguas tonais e com uma variedade de contextos linguísticos (combinações de vogais, consoantes e tons lexicais) abrangente.

1.2 Objetivos

O objetivo deste trabalho é **quantificar a contribuição da informação visual na percepção de tons lexicais**. Este trabalho busca aplicar uma abordagem que corrobore de forma quantitativa as evidências da multimodalidade do tom lexical apresentadas por outros trabalhos (BURNHAM; LAU et al., 2001; BURNHAM; CIOCCA; STOKES, 2001; MIXDORFF; HU; BURNHAM, 2005; CHEN; MASSARO, 2008; SMITH; BURNHAM, 2012; BURNHAM; KASISOPA et al., 2015; GARG et al., 2019). Dentro desse escopo, os objetivos específicos do trabalho são os seguintes:

- determinar a capacidade da informação visual (sinais de movimento da face e da cabeça) de classificar tons lexicais e comparar essa capacidade com a do sinal acústico;
- determinar quais sinais de movimento (face ou cabeça) possuem maior capacidade de classificação de tons lexicais;
- verificar se existem, em línguas específicas, tons lexicais que são mais fáceis do que outros de serem discriminados a partir da informação visual (sinais de movimento da face e da cabeça);
- determinar quais componentes dos sinais de movimento da face e da cabeça são mais relevantes para a classificação de tons lexicais.

Para tentar responder a essas perguntas, este trabalho propõe um novo método quantitativo baseado em técnicas de classificação estatística. Neste método, dados de movimento da face e da cabeça são usados para treinar classificadores cujo desempenho é então avaliado em tarefas de classificação de tons lexicais. Essa é uma abordagem quantitativa baseada na parametrização da produção dos tons lexicais que é capaz de extrair informações importantes, como as componentes visuais mais relevantes na separação dos tons. Neste trabalho, isso será realizado de forma mais direta do que em outros trabalhos que utilizaram outras abordagens quantitativas, como ajuste de modelos lineares mistos, análise de variância ou análise de componentes principais (BURNHAM; LAU

et al., 2001; BURNHAM; CIOCCA; STOKES, 2001; MIXDORFF; HU; BURNHAM, 2005; BURNHAM; KASISOPA et al., 2015; BURNHAM; LI et al., 2019). Utilizamos essa abordagem, pois a percepção de tons lexicais pode ser encarada como um problema de classificação. Seja uma situação de comunicação por meio de fala numa língua tonal: com base numa entrada, que pode ser acústica, visual ou audiovisual, cada interlocutor interpreta as palavras ditas como portadoras de um dos possíveis tons lexicais (6 no cantonês, 4 no mandarim ou 5 no tailandês, por exemplo). Pode-se então fazer a seguinte analogia: as palavras ditas são nossa base de dados e o interlocutor que as interpreta é o classificador estatístico.

Este trabalho é contextualizado na literatura da área por procurar preencher as seguintes lacunas deixadas por outros estudos: 1) utiliza uma metodologia nova para este problema, baseada em classificação estatística a partir de informações audiovisuais da produção de tons lexicais; 2) trabalha com técnicas de classificação estatística que são interpretáveis, de modo que seja possível saber quais são os movimentos específicos que mais influenciam na classificação dos tons lexicais; e 3) trabalha com dados acústicos e visuais produzidos em diversos contextos linguísticos por falantes nativas de três línguas tonais distintas: cantonês, mandarim e tailandês.

1.3 Estrutura do trabalho

O restante do trabalho está estruturado como descrito a seguir. A base teórica necessária para a realização e entendimento do trabalho é dividida em dois capítulos, cobrindo os temas produção da fala (Capítulo 2) e línguas tonais (Capítulo 3). A seguir, a aquisição e o processamento da base de dados utilizada no trabalho são detalhadas no Capítulo 4 e as técnicas de classificação utilizadas são descritos no Capítulo 5. Após isso, no Capítulo 6 são apresentados e discutidos os resultados. A conclusão do trabalho é apresentada, por fim, no Capítulo 7.

2 Produção da fala

A produção da fala é um fenômeno complexo dependente de uma série de conceitos para ser compreendido. Deste modo, este capítulo busca detalhar o processo de produção da fala, que é parte da Cadeia da Fala (DENES; PINSON, 2015), tratando especificamente dos níveis fisiológico e acústico, desde o ar saindo dos pulmões do falante até a propagação das ondas sonoras irradiadas pelas cavidades oral e nasal.

2.1 Aparelho fonador

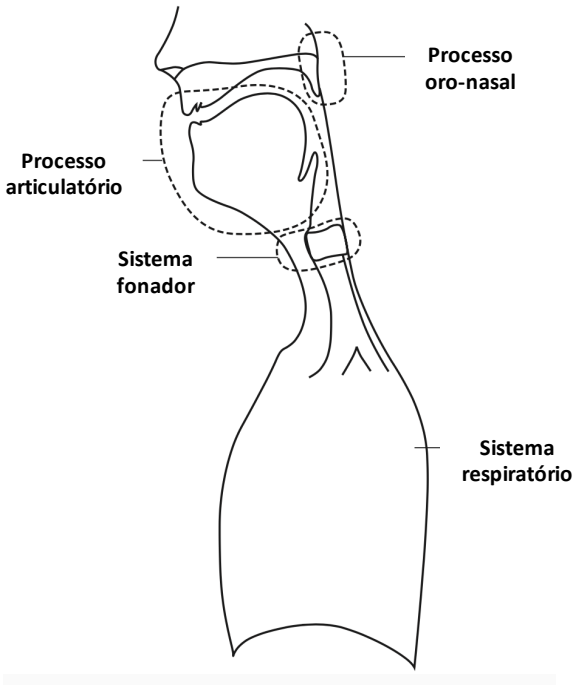
O corpo humano não é dotado de um sistema específico para a produção da fala do mesmo modo que é dotado de sistemas específicos para respiração, digestão e circulação. A fala é produzida pela ação coordenada dos sistemas respiratório, fonatório e articulatório, que em conjunto formam o aparelho fonador (SILVA et al., 2019). Enquanto os sistemas respiratório e fonatório são responsáveis pelo processo de geração de som a partir da vibração das pregas vocais, chamado de fonação, o sistema articulatório confere características específicas a esse som de acordo com os articuladores empregados. A Figura 2 detalha o aparelho fonador e seus componentes, dividindo o sistema articulatório no processo que ocorre na cavidade oral (articulatório) e no processo de acoplamento com a cavidade nasal (oro-nasal).

O sistema respiratório é responsável por controlar a entrada e saída de ar no trato vocal. Por sua vez, o sistema fonatório tem a capacidade de obstruir ou permitir a passagem desse ar por meio da glote, que é o espaço entre as pregas vocais. O movimento de abertura e fechamento da glote corresponde a um ciclo glotal, e a repetição de ciclos glotais caracteriza a fonação (SILVA et al., 2019), que será descrita com maior detalhe na Seção 2.1.3.

Os sons da fala podem ser classificados com relação à presença ou não de fonação em vozeados e não-vozeados. Enquanto que nos sons vozeados há presença de fonação, ou de sucessivos ciclos glotais, nos sons não-vozeados não há fonação, pois as pregas vocais permanecem afastadas uma da outra (SILVA et al., 2019). A Figura 3 exemplifica as configurações da glote durante a produção de sons vozeados, à esquerda, e de sons não-vozeados, à direita.

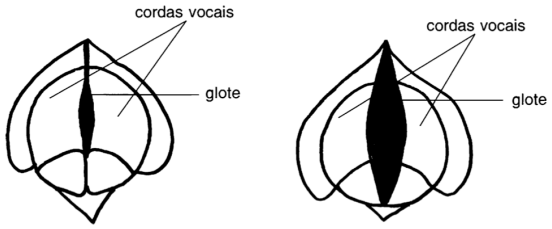
O sistema articulatório, que compreende os articuladores passivos e ativos, é então excitado pelos pulsos de ar provenientes da fonação. Os articuladores passivos, que não se movimentam, são o lábio superior, os dentes, os alvéolos, o palato duro, o palato mole e a úvula. Os articuladores ativos, que se movimentam em direção a articuladores

Figura 2 – Detalhamento do sistema fonador.



Fonte: Adaptado de (LADEFOGED; JOHNSON, 2010, Cap. 1)

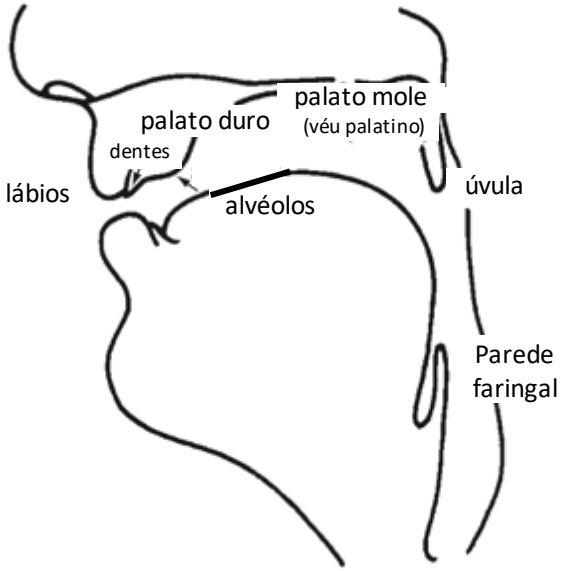
Figura 3 – Detalhamento da glote durante a produção de sons vozeados e não-vozeados.



Fonte: (SILVA, 2003, p. 28)

passivos, são o lábio inferior, a língua, o véu palatino e as pregas vocais (SILVA et al., 2019). Diferentes tipos de aproximação entre os articuladores produzem diferentes sons, que se dividem basicamente em vogais e consoantes. As Figuras 4 e 5 ilustram a posição dos articuladores que compõem o sistema articulatório na parte superior e inferior da cavidade oral. Enquanto na produção de vogais há pouca aproximação entre articuladores, na produção de consoantes os articuladores se aproximam bastante, chegando até a se encostarem. Cada língua utiliza um subconjunto do total de sons da fala conhecidos, descritos em (INTERNATIONAL PHONETIC ASSOCIATION, 2015), ou seja, em uma dada língua apenas algumas vogais e algumas consoantes, com seus respectivos parâmetros

Figura 4 – Posições dos articuladores na parte superior da cavidade oral.



Fonte: Adaptado de (LADEFOGED; JOHNSON, 2010, Cap. 1)

articulatórios, são utilizadas. Nas próximas seções serão detalhados os processos de produção específicos de vogais e consoantes, assim como o processo de fonação.

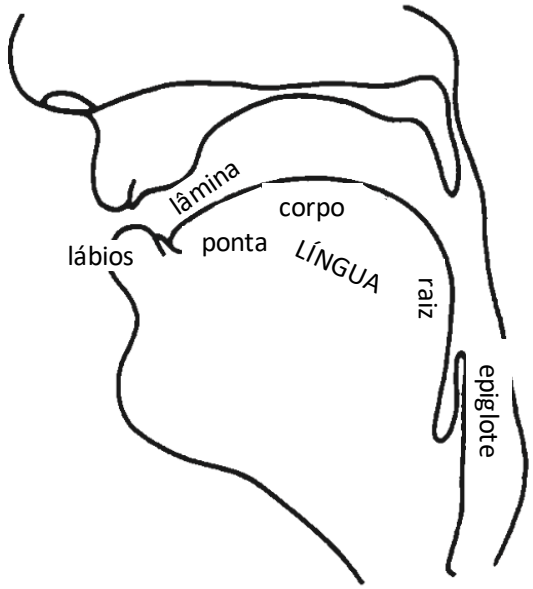
2.1.1 A produção de vogais

As vogais são sons produzidos com uma resistência mínima à passagem do ar pelo trato vocal e são tipicamente vozeadas. Os parâmetros articulatórios que caracterizam os sons vocálicos são os seguintes, de acordo com (LADEFOGED; JOHNSON, 2010, Cap. 1):

- Altura da língua ou abertura/fechamento da mandíbula;
- Avanço/recuo da língua;
- Arredondamento/estreitamento dos lábios;
- Abertura/fechamento do véu palatino.

Anatomicamente, a altura da língua está relacionada à abertura e ao fechamento da mandíbula: quanto mais próxima do palato a língua estiver, mais alta ela está e, consequentemente, mais fechada está a mandíbula. Por outro lado, quando mais afastada do palato a língua estiver, mais baixa ela está e mais aberta está a mandíbula. No português brasileiro, um exemplo de vogal alta é [i], enquanto um exemplo de vogal baixa é [a].

Figura 5 – Posições dos articuladores na parte inferior da cavidade oral.



Fonte: Adaptado de (LADEFOGED; JOHNSON, 2010, Cap. 1)

Outro parâmetro articulatório importante para as vogais é o avanço/recuo da língua. Vogais podem ser classificadas como anteriores, quando a língua está mais próxima da região alveolar, posteriores, quando a língua está mais próxima da região do véu palatino, e centrais, quando a língua está num ponto intermediário. No português brasileiro, [i] é um exemplo de vogal anterior, [a] é um exemplo de vogal central e [u] é um exemplo de vogal posterior.

Os lábios também participam da articulação das vogais, arredondando-se, ou estreitando-se. As vogais podem ser classificadas, portanto, como arredondadas, quando os lábios estão aproximados e projetados para frente, e não-arredondadas, quando os lábios estão estirados. No português brasileiro, [a] é um exemplo de vogal não-arredondada e [u] um exemplo de vogal arredondada.

Por fim, a abertura ou fechamento do véu palatino é também um parâmetro articulatório relevante para as vogais. Se o véu palatino estiver levantado, não há acoplamento entre as cavidades oral e nasal e o ar só escapa pela boca, caracterizando o som vocálico como oral. Por outro lado, se o véu palatino estiver abaixado, acoplando as duas cavidades, o ar escapa tanto pela boca quanto pelo nariz e o som vocálico gerado é nasal. Um exemplo de vogal nasal é o primeiro [a] na palavra manta, que se contrasta com o primeiro [a] da palavra mata, que não é nasal.

Os quatro parâmetros articulatórios descritos acima descrevem especificamente os sons vocálicos e, apesar de possuírem natureza contínua entre seus extremos, são

classificados em categorias discretas (SILVA et al., 2019). Deste modo, as vogais são nomeadas com as categorias discretas que melhor descrevem o estados desses parâmetros articulatórios na mesma ordem em que eles foram apresentados neste trabalho. Um exemplo é a vogal [i], cuja nomenclatura é vogal alta (ou fechada) anterior não-arredondada. Subentende-se que a vogal é oral se não houver alusão à nasalidade em sua nomenclatura.

Os sons vocálicos ainda podem ser não-vozeados, quando perdem parte da capacidade de vozeamento, e também, quando vozeados, classificados como monotongos ou ditongos (SILVA et al., 2019). Este trabalho, contudo, não se ocupará com o detalhamento dessas características, atendo-se aos quatro parâmetros articulatórios básicos. Mais exemplos de vogais no português brasileiro estão disponíveis em (CRISTÓFARO-SILVA; YEHIA, 2009).

2.1.2 A produção de consoantes

Diferentemente das vogais, as consoantes são sons produzidos com resistência significativa à passagem de ar no trato vocal. Devido a essa natureza, os parâmetros articulatórios que caracterizam os sons consonantais são os seguintes, segundo (LADEFOGED; JOHNSON, 2010, Cap. 1):

- Modo de articulação;
- Ponto de articulação;
- Vozeamento.

O modo de articulação se refere ao tipo de resistência à passagem de ar que ocorre durante a produção da consoante, que pode ser desde um bloqueio total até um bloqueio mínimo. As consoantes podem ser classificadas de acordo com o modo de articulação em categorias como oclusivas, nasais, fricativas, africadas, tepes, vibrantes, aproximantes retroflexas e laterais. O parágrafo seguinte descreve os modos de articulação e traz exemplos do português brasileiro.

A produção de consoantes oclusivas, também chamadas de plosivas, envolve o bloqueio total da passagem do ar no trato vocal por meio do encontro de dois articuladores ([p] em pá e [t] em tom). As consoantes nasais também são caracterizadas pelo bloqueio total da passagem de ar no trato vocal na cavidade oral, mas contam com a abertura do véu palatino, que permite que o ar escape pelas narinas, ao contrário das plosivas, que são articuladas com o véu palatino fechado ([m] em mau e [n] em neve). Outro modo de articulação é o das consoantes fricativas, caracterizadas por uma grande aproximação entre dois articuladores que causa alta resistência à passagem de ar ([f] em fé e [s] em selo). Com a proximidade entre os articuladores, a área pela qual o ar deve fluir é muito

pequena, gerando fricção entre suas partículas, o que soa como um ruído. As consoantes africadas, por sua vez, são articuladas por meio de uma oclusão seguida de fricção: há um bloqueio total da passagem de ar que é então desfeito, permitindo o fluxo de ar por uma área muito pequena entre dois articuladores ([dʒ] em dia). Combina-se, portanto, os modos de articulação das oclusivas e das fricativas, em sequência. Há também consoantes articuladas como tepes, termo que vem da língua inglesa, *tap*, e que significa uma batida repentina e breve (SILVA et al., 2019). Os tepes são articulados de forma a bloquear rapidamente a passagem de ar pelo trato vocal, permitindo o prosseguimento do fluxo em seguida ([ɾ] em prova). De modo semelhante aos tepes, as consoantes vibrantes são articuladas por meio de bloqueio rápido à passagem de ar no trato vocal ([r̆] em rato). A diferença é que na produção de vibrantes há uma série de bloqueios sucessivos que soam como uma vibração. A articulação das consoantes aproximantes retroflexas, por outro lado, é caracterizada pela curvatura da ponta da língua sobre seu corpo, gerando resistência parcial à passagem do ar ([ɹ] em mar). Por fim, as consoantes cuja produção envolve menor resistência à passagem de ar são as laterais. Ela se caracterizam pela formação simultânea de um bloqueio (central) e canais de passagem (laterais) pela língua, fazendo com o ar se propague lateralmente ([l] em lã e [ʎ] em malha).

O segundo parâmetro articulatório que caracteriza as consoantes é o ponto de articulação, que indica quais são os articuladores passivos e ativos envolvidos na produção de cada som. Em termos de ponto de articulação, as consoantes podem também ser classificadas em categorias como alveolares, alveolo-palatais, bilabiais, dento-alveolares, glotais, labiodentais, labiovelares, palatais e velares.

O terceiro e último parâmetro articulatório característico das consoantes é o vozeamento, que reflete a ação das pregas vocais. Se há ação das pregas vocais durante a produção da consoante, ela é vozeada. Caso contrário, ela é não-vozeada. As consoantes [p] e [b], por exemplo, são ambas oclusivas bilabiais, mas [p] é não-vozeada e [b] é vozeada.

Além disso, uma consoante pode ser aspirada, caso haja a presença de um período de não-vozeamento durante e após sua articulação. Exemplos de consoantes aspiradas são as oclusivas [k^h], [p^h] e [t^h], presentes, por exemplo, em inglês nas palavras *kiss*, *pat* e *tap* (LADEFOGED; JOHNSON, 2010, Cap. 3).

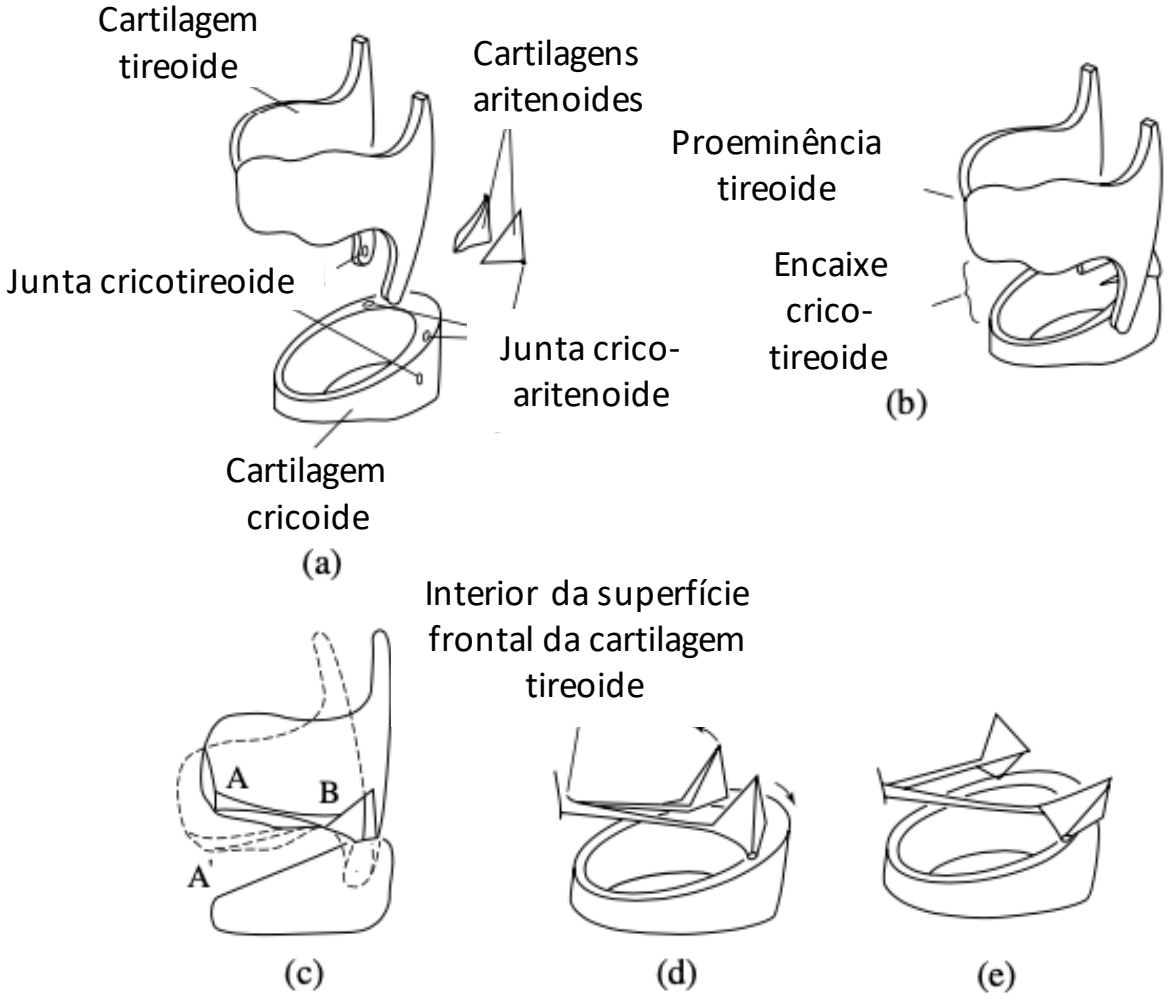
Deste modo, cada consoante é nomeada de acordo com os parâmetros articulatórios que a caracterizam na ordem em que foram detalhados acima. Por exemplo, a consoante [p] é uma oclusiva bilabial vozeada empregada no português brasileiro em palavras como **p**i**p**oca e **p**lanta e no inglês em palavras como **p**ie. Outro exemplo é a consoante [z], uma fricativa alveolar vozeada empregada no português brasileiro em palavras como **z**ero e a**z**eite e no inglês em palavras como **z**oo. Mais exemplos de consoantes no português brasileiro estão disponíveis em (CRISTÓFARO-SILVA; YEHIA, 2009).

2.1.3 Fonação

Nos seres humanos, a laringe possui quatro funções básicas: 1) a proteção das vias aéreas, particularmente importante durante a deglutição; 2) a fixação do tronco para movimentos das extremidades superiores; 3) a abertura das vias aéreas para a respiração; e, finalmente, 4) a fonação (HIROSE, 2010). Neste trabalho apenas a fonação será descrita em detalhes.

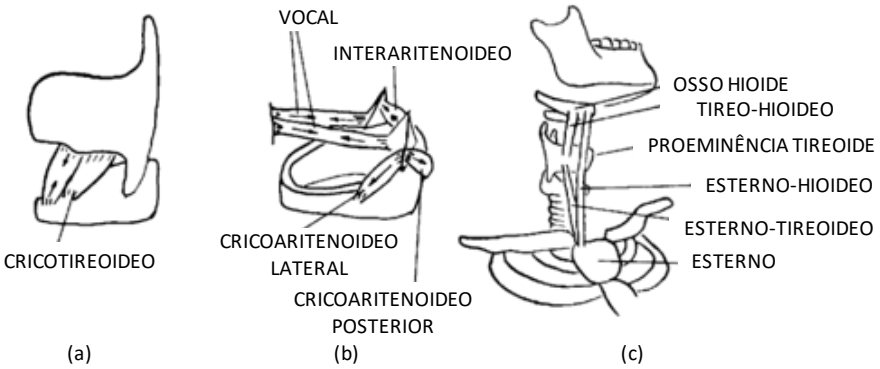
A laringe é capaz de desempenhar as funções acima por meio de sua estrutura composta por cartilagens, músculos e outros tecidos. A estrutura principal da laringe é composta por quatro cartilagens principais, ilustradas pela Figura 6. A cartilagem tireoide, um anel aberto, é localizada acima da cartilagem cricoide, um anel fechado. Ainda há duas cartilagens menores chamadas de aritenoides, localizadas acima da parte posterior da cricoide (OHALA, 1978). A disposição das quatro cartilagens pode ser visualizada nos desenhos (a) e (b) da Figura 6. As pregas vocais são músculos conectados às cartilagens tireoide e aritenoides, visualizadas nos desenhos (d) e (e) da Figura 6. O espaço entre as pregas vocais é chamado de glote e permite a passagem de ar dos pulmões ao trato vocal. A rotação das cartilagens aritenoides pode aproximar ou afastar as pregas vocais, deste modo fechando ou abrindo a glote (YIP, M., 2002, Seção 1.2.1). Além disso, as cartilagens estão conectadas entre si e sua articulação tem papel crucial na fonação.

Figura 6 – Ilustrações da laringe sob diversas perspectivas. (a) vista explodida das cartilagens; (b) cartilagens encaixadas normalmente; (c) modo de rotação das cartilagens tireoide e cricoide que alongam as pregas vocais de AB a $A'B$. É um movimento de báscula que tensiona as pregas vocais; (d) posição aduzida das pregas vocais com as cartilagens aritenoides inclinada para dentro; (e) posição abduzida das pregas vocais com as cartilagens aritenoides inclinadas para fora. Em (d) e (e) são representadas com clareza as cartilagens aritenoides, que controlam a massa vibrante das pregas vocais.



Fonte: Adaptado de (OHALA, 1978).

Figura 7 – Representações esquemáticas dos músculos da laringe e seus movimentos.



Fonte: Adaptado de (OHALA, 1978)

Os movimentos das articulações cricotireoidea e cricoaritenoidea são controlados pelos músculos da laringe, ilustrados na Figura 7. (HIROSE, 2010) descreve o funcionamento dos músculos da laringe: a contração do músculo cricotireoideo (CT) alonga as pregas vocais. Existem também músculos abdutores e adutores, que movimentam a cartilagem aritenoide: o músculo cricoaritenoideo posterior (PCA) é o único abdutor, enquanto que os músculos interaritenoideo (IA), cricoaritenoideo lateral (LCA) e tireoaritenoideo (TA) são adutores. O músculo vocal (VOC), que é parte do músculo TA, controla a massa e a rigidez das pregas vocais. Todos os músculos descritos acima são ilustrados nos desenhos (a) e (b) da Figura 7. Por outro lado, os músculos ilustrados no desenho (c) da Figura 7 são responsáveis pelo suporte da laringe, podendo também elevá-la ou abaixá-la como um todo.

A vibração das pregas vocais durante a fonação pode ser descrita de acordo com a teoria mioelástica-aerodinâmica. (HIROSE, 2010) descreve as seguintes etapas envolvidas na fonação, de acordo com essa teoria: um ciclo de vibração das pregas vocais se inicia com a aproximação das duas pregas por meio da ativação de músculos adutores. Em seguida, o ar é forçado dos pulmões até o trato vocal, o que faz com que as pregas vocais se juntem por meio do efeito combinado da lei de Bernoulli e da elasticidade dos tecidos. Com o fechamento da glote e a manutenção do fluxo de ar saindo dos pulmões, a pressão na cavidade subglótica aumenta até forçar a reabertura da glote, que permite a passagem de um pulso de ar até o trato vocal. Com essa passagem de ar, a pressão na cavidade subglótica cai e as pregas vocais voltam a suas posições iniciais devido à elasticidade de seus tecidos e à força resultante da diferença de pressão entre as cavidades sub e supraglóticas. Um novo ciclo de vibração pode então iniciar (HIROSE, 2010).

O processo de fonação foi descrito acima de forma básica. Existem ainda ajustes que podem ser feitos nesse processo de modo a causar variações de frequência fundamental

(F0). Sabe-se que alterações no comprimento, na rigidez e na espessura das pregas vocais causam variações de F0 (HIROSE, 2010). Por um lado, a contração do músculo CT alonga as pregas vocais, aumentando sua rigidez, o que resulta na produção de sons com valores mais altos de F0. Por outro lado, o relaxamento do músculo CT aliado à contração do músculo TA, aumenta a espessura das pregas vogais, resultando em sons com valores mais baixos de F0 (HIROSE, 2010; YIP, M., 2002, Seção 1.2.1) Outro mecanismo que pode controlar os valores de F0 dos sons produzidos na fonação é o abaixamento da laringe, que estica as pregas vocais, diminuindo sua espessura, resultando em valores menores de F0 (OHALA, 1978).

2.2 Propriedades acústicas da fala

A seção anterior detalhou o processo de produção da fala com foco em seus aspectos fisiológicos. Como resultado da ação do aparelho fonador são produzidos sons que devem se propagar até um ouvinte para que haja comunicação. Esta seção trata dessa propagação, caracterizada como o nível acústico da Cadeia da Fala (DENES; PINSON, 2015).

A acústica como ciência pode ser descrita como a geração, transmissão e recepção de energia na forma de ondas vibracionais na matéria (KINSLER et al., 2000, Cap. 1). Quando as moléculas de um fluido, como o ar, ou de um sólido, como os ossos humanos, são deslocadas de suas posições originais, surge uma força interna de restauração. Essa força, em conjunto com a inércia do sistema, possibilita à matéria vibrar de forma oscilatória, gerando e transmitindo ondas acústicas (KINSLER et al., 2000, Cap. 1). Os sons gerados pelo aparelho fonador são um fenômeno acústico cujas propriedades serão detalhadas a seguir. Primeiramente será detalhado um modelo acústico de produção da fala e, em seguida, os aspectos físicos da onda sonora.

2.2.1 Teoria Acústica de Produção da Fala

Um modelo muito difundido da produção de fala é baseado na Teoria Acústica de Produção da Fala (FANT, 1960). Essa teoria estabelece o modelo Fonte-Filtro, segundo o qual a onda sonora é a resposta dos sistemas de filtro do trato vocal a uma ou mais fontes sonoras e a fala pode, portanto, ser especificada em termos das características do filtro e da fonte. Simbolizando a fonte por E e o filtro por V, tem-se o som da fala como o produto $S = E * V$, sendo que cada uma dessas variáveis é dependente do tempo e da frequência.

Pode-se relacionar os termos fonéticos fonação e articulação aos termos fonte e filtro, respectivamente. Isso indica que para a Teoria Acústica de Produção da Fala a fonação é um fenômeno separado e independente da articulação: enquanto a primeira gera o som, a segunda modela suas qualidades fonéticas (FANT, 1960).

Prosseguindo no detalhamento desse modelo, o trato vocal é interpretado como um tubo acústico fechado num lado e aberto no outro cujas terminações são a glote (terminação fechada) e as cavidades oral e nasal (terminação aberta). Tubos desse tipo (de comprimento L) são ressonadores de $\frac{1}{4}$ de onda, ou seja, facilitam a propagação de ondas com comprimento λ específico que respeite a relação $L = \frac{n\lambda}{4}, n = 1, 3, 5, \ldots$. As frequências dessas ondas são as frequências de ressonância do trato vocal que, por meio de excitação da fonte, geram as frequências formantes.

É conveniente definir a fonte sonora como o fluxo de ar pulsante através da glote, representado no domínio do tempo por uma série periódica de pulsos e no domínio da frequência por um espectro harmônico. As características do espectro da fonte são amplificadas ou atenuadas de acordo com o espectro do filtro (o trato vocal) (FANT, 1960):

$$|S(j\Omega)| = |E(j\Omega)| * |V(j\Omega)| \tag{2.1}$$

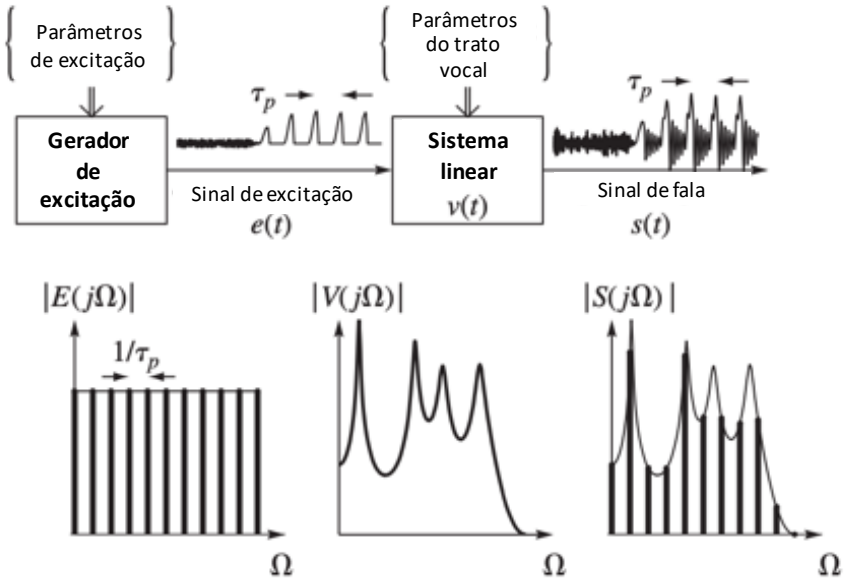
O sinal $|S(j\Omega)|$ é o espectro do som da fala e seus picos são chamados de frequências formantes. A posição dos articuladores do falante moldam as características do filtro, o que reflete diretamente na posição das frequências formantes. Como cada som da fala possui parâmetros articulatórios característicos, cada som da fala também possui um conjunto de frequências formantes característico.

A Figura 8 ilustra o modelo Fonte-Filtro proposto em (FANT, 1960). Na parte superior está a representação temporal do modelo: o sinal de excitação $e(t)$, a fonte, é a entrada de um sistema $v(t)$, o filtro, cuja saída é o sinal de fala $s(t)$. Na parte inferior está a representação espectral do modelo, no domínio da frequência: os vários picos de $|E(j\Omega)|$, harmônicos provenientes da fonação, são modulados pelo espectro do filtro $|V(j\Omega)|$ resultando no espectro da fala $|S(j\Omega)|$, cujos picos são chamados de frequências formantes e dependentes de $|V(j\Omega)|$, que por sua vez depende da configuração do trato vocal.

Uma interessante relação entre os domínios do tempo e da frequência é o espaçamento entre os picos dos sinais $e(t)$ e $|E(j\Omega)|$. No domínio do tempo, os picos de $e(t)$ são espaçados entre si pelo período τ_p, que é uma medida de tempo, indicando que seu gráfico apresenta variações de amplitude ao longo do tempo. Por outro lado, no domínio da frequência, os picos de $|E(j\Omega)|$ são espaçados entre si pelo inverso do período, que é uma medida de frequência, indicando que seu gráfico apresenta a amplitude que cada componente de cada frequência específica possui. Um sinal periódico simples, como uma senoide, é representado no domínio da frequência por um impulso no valor de sua frequência. Um sinal periódico complexo, como alguns sons da fala, por sua vez, é representado no domínio da frequência por uma série de picos, cada um representando um sinal periódico simples que o compõe.

Em resumo, a fala, segundo a Teoria Acústica de Produção da Fala (FANT, 1960),

Figura 8 – Modelo Fonte-Filtro linear de produção da fala com representações temporais e espectrais da fonte, do filtro e do sinal resultante.



Fonte: Adaptado de (RABINER; SCHAFER, 2011, Seção 3.2.3).

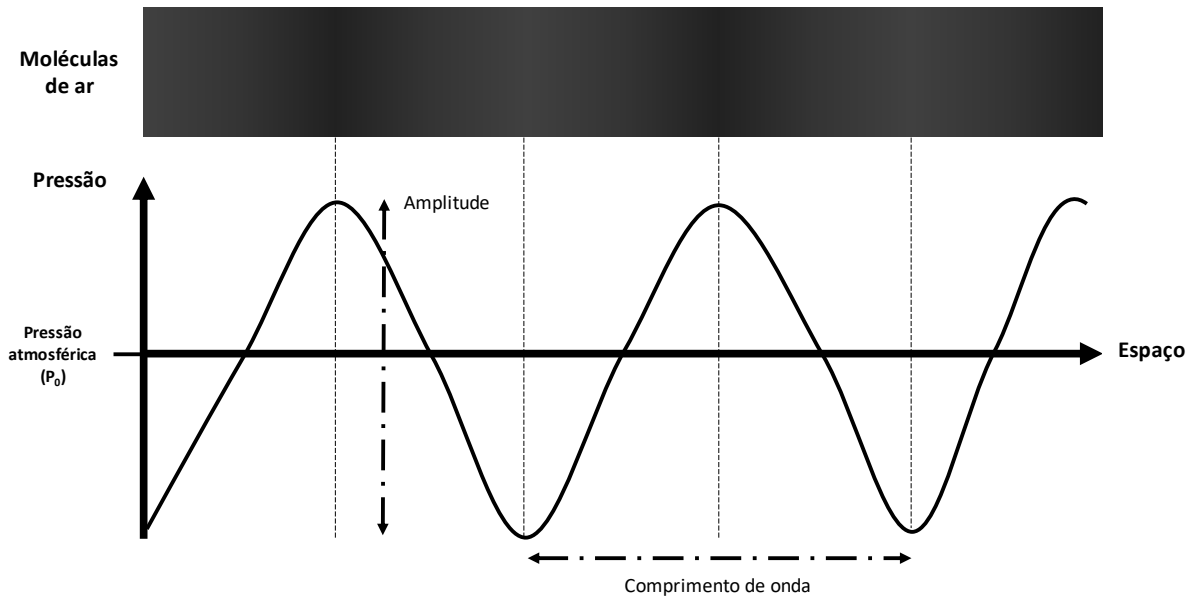
é a resposta de um filtro (trato vocal) à excitação de uma fonte (fluxo de ar através da glote) e é transmitida através do ar.

2.2.2 A onda sonora

Fisicamente, o fenômeno da onda sonora é resultante do processo de produção de fala e se propaga através do ar. Um fenômeno oscilatório é relacionado a movimentos em torno de um ponto de equilíbrio, e a propagação da fala como onda sonora acontece pela aproximação e pelo afastamento das moléculas de ar em torno da posição de repouso, gerando sucessivas zonas de alta pressão (onde há maior concentração de moléculas) e de baixa pressão (onde há menor concentração de moléculas) (SILVA et al., 2019). A Figura 9 traz, em sua porção superior, a organização das moléculas de ar que transmitem uma onda sonora e, em sua porção inferior, a representação dos níveis de pressão em cada ponto do ar. As zonas onde há maior concentração de moléculas de ar são representadas na porção superior como mais escuras e na porção inferior como os picos positivos da senoide. As zonas onde há menor concentração de moléculas de ar são, por outro lado, representadas na porção superior como mais claras e na porção inferior como os picos negativos da senoide.

Uma fonte sonora é um corpo que, ao vibrar, transmite oscilações às moléculas de ar adjacentes, gerando as variações de pressão citadas acima. No caso da fala, a fonte

Figura 9 – Representação da onda sonora.



Fonte: o autor.

sonora são as pregas vocais, que vibram gerando um fluxo de ar pulsante que passa através da glote. Os parâmetros físicos que caracterizam as ondas sonoras são os seguintes:

- Frequência;
- Amplitude;
- Fase.

A frequência da onda sonora da fala corresponde ao número de ciclos glotais completados por segundo e é medida em hertz [Hz]. Os seres humanos têm, idealmente, capacidade de perceber sons na faixa de frequência de 20Hz a 20kHz, existindo ainda sons mais graves, abaixo de 20Hz (infrassons), e mais agudos, acima de 20kHz (ultrassons), que não são perceptíveis aos humanos (SILVA et al., 2019).

Um ciclo glotal corresponde a um movimento de abertura seguido por um movimento de fechamento da glote, inicialmente fechada. O tempo de duração de um ciclo é chamado de período de vibração (T) e é inversamente proporcional à frequência de vibração, também chamada de frequência fundamental (F0). Sua relação é expressa pela seguinte equação:

$$T = \frac{1}{F0} \tag{2.2}$$

Em termos perceptivos, a frequência fundamental se relaciona com a sensação de um som ser grave ou agudo. Quanto menor o valor de F0, o som será percebido como mais

grave. A fonação de cada falante pode produzir sons dentro de uma faixa de F0 que varia de acordo com propriedades do corpo vibrante (no caso as pregas vocais) como massa, tensão, volume e forma (SILVA et al., 2019). De forma geral, homens possuem F0 menor que o de mulheres, que, por sua vez, possuem F0 menor que o de crianças.

A amplitude de uma onda sonora é relacionada diretamente à intensidade acústica, que é definida como a variação da energia acústica em um intervalo de tempo através de uma área, cuja unidade de medida é o watt por metro quadrado [W/m^2] (SILVA et al., 2019). Um som produzido com muita intensidade é um som de alta amplitude enquanto que um som produzido com pouca intensidade é um som de baixa amplitude. De modo análogo à frequência, a amplitude possui valores máximos e mínimos passíveis de percepção pelos seres humanos. O menor valor de intensidade percebido por humanos é chamado de limiar da audição e equivale a 10^{-12} W/m^2 e o maior valor capaz de ser percebido sem causar dor é chamado de limiar da dor e equivale a 10^{13} W/m^2. Devido à grande variabilidade dos valores percebidos pelos seres humanos, as medidas de intensidade são geralmente feitas em escala logarítmica e representadas pela grandeza nível de intensidade (KINSLER et al., 2000, Cap. 11):

$$I_L = 10 \log \left(\frac{I}{I_{ref}} \right) \tag{2.3}$$

Na equação acima, I_L é o nível de intensidade medido em decibéis [dB], resultante do produto do escalar 10 com o logaritmo da razão entre a intensidade de interesse e a intensidade de referência é equivalente à intensidade do limiar da audição $I_{ref} = 10^{-12}$ W/m^2.

Por fim, a fase corresponde ao ponto do ciclo em que a onda se encontra em determinado momento. A fase é importante na combinação de ondas sonoras diferentes, pois influencia a relação entre elas. A fala, por exemplo, é fruto da combinação de ondas sonoras com diferentes frequências, amplitudes e fases (SILVA et al., 2019).

Esses três parâmetros físicos são suficientes para descrever qualquer onda sonora simples. Contudo, as ondas sonoras da fala são complexas e podem ser classificadas também como periódicas ou aperiódicas, considerando a presença ou não de ciclos de repetição. Dentre as ondas sonoras periódicas existem as simples e as complexas, a depender de quantos períodos diferentes a compõem, e dentre as ondas sonoras aperiódicas existem as transientes e as contínuas, a depender de sua duração (SILVA et al., 2019).

Uma onda sonora periódica simples é caracterizada por uma única frequência. Por outro lado, uma onda sonora periódica complexa é um somatório de ondas periódicas simples com frequências, amplitudes e fases distintas. Essa soma ocorre pelo Princípio da Superposição de Ondas, que afirma que se duas ou mais ondas passam por um dado ponto, em um determinado instante, a amplitude resultante será a soma algébrica da amplitude

de cada uma das ondas (SILVA et al., 2019). Além disso, as frequências das diversas ondas sonoras periódicas simples que compõem a onda sonora periódica complexa são chamadas de harmônicos, definidos como os múltiplos inteiros da frequência mais baixa, chamada de frequência fundamental (F0) (SILVA et al., 2019).

As ondas sonoras aperiódicas, por sua vez, não possuem ciclos glotais em intervalos regulares de tempo. Ondas aperiódicas transientes são aquelas que ocorrem numa curta extensão de tempo como, por exemplo, as consoantes oclusivas. Por outro lado, as ondas aperiódicas contínuas ocorrem numa extensão maior de tempo como, por exemplo, as consoantes fricativas (SILVA et al., 2019).

As ondas sonoras da fala são compostas por trechos de relativa periodicidade e por trechos de aperiodicidade. A periodicidade é relativizada devido à duração dos sucessivos ciclos glotais, que possuem valores numericamente muito próximos, mas ainda diferentes. Deste modo, as ondas sonoras da fala são chamadas de quasi-periódicas (SILVA et al., 2019).

3 Línguas tonais

O capítulo anterior tratou da produção da fala de um modo geral, sem contemplar línguas específicas. Este capitulo, contudo, tratará de detalhar as línguas tonais, discutindo suas especificidades fisiológicas, acústicas e linguísticas. Além disso, serão descritas as estruturas fonológicas das línguas tonais cantonês, mandarim e tailandês, objeto de estudo deste trabalho. Ao fim do capítulo, procura-se justificar a pergunta norteadora deste trabalho, afastando-a da procura por uma relação espúria e aproximando-a do desenvolvimento adicional a uma série de estudos que caminham na mesma direção.

3.1 O que é uma língua tonal

Os parâmetros articulatórios descritos no Capítulo 2, como a abertura da mandíbula para as vogais e o ponto de articulação para as consoantes, são comuns a todas as línguas. Portanto, nenhuma língua é caracterizada como sendo uma língua de abertura da mandíbula ou de ponto de articulação. Por outro lado, o tom está presente apenas em um subconjunto das línguas, que são chamadas de línguas tonais e são cerca de 70% das línguas do mundo (YIP, M., 2002).

Todas as línguas do mundo utilizam variações de frequência fundamental (F0) como parte de seus sistemas sonoros e o que as difere são as funções dessas variações (MADDIESON, 2013). Os dois usos mais comuns da variação de F0 nas línguas são a entoação e o tom, que possuem diferentes domínios de aplicação. Entoação é o termo que descreve uma variação de F0 que se dá ao longo de um enunciado ou de parte de um enunciado, agregando a ele outros sentidos. Por exemplo, a entoação pode diferenciar uma afirmação de uma pergunta, indicar se o falante deseja continuar falando ou não e também contrastar a apresentação de informações novas e antigas (MADDIESON, 2013). O tom, por sua vez, é o uso de padrões de variação de F0 em palavras isoladas, alterando seu significado ou sua função gramatical. Padrões comuns de variação de F0 em tons são, por exemplo, um F0 constante em um determinado valor de frequência, chamado de nível, ou uma variação crescente ou decrescente do valor de F0 ao longo da palavra, chamada de contorno (MADDIESON, 2013).

Línguas tonais são, portanto, aquelas que fazem uso do tom na distinção semântica das palavras. Com essa função, o tom passa a se chamar tom lexical (YIP, M., 2002). As regiões do mundo com a maior concentração de línguas tonais são a África, o leste e sudeste da Ásia, o Pacífico e as Américas (YIP, M., 2002), sendo exemplos o yorubá (WARD, 1952), falado em países como a Nigéria, Benin e Togo, o tikuna (SOARES, 1986), falado na região amazônica de países como o Brasil e a Colômbia, e as línguas chinesas (YIP, M. J.,

1980), entre elas o mandarim e o cantonês. Mapas com a localização geográfica de falantes de línguas tonais no mundo podem ser encontrados em (YIP, M., 2002) e (MADDIESON, 2013). Para ilustrar o funcionamento do tom lexical nas línguas tonais, pode-se citar o cantonês: ao falar a sílaba [yau], ela pode significar preocupação, se pronunciada com um tom de nível alto (F0 alto e constante), ou óleo, se pronunciada com um tom de nível baixo (F0 baixo e constante) (YIP, M., 2002, Seção 1.1).

Dentre as línguas que não são tonais, existem as línguas acentuais, caracterizadas pela presença de sílabas acentuadas. Acusticamente, o acento é caracterizado principalmente por sua duração e intensidade: vogais e consoantes em sílabas acentuadas tendem a ser mais longas do que em sílabas não acentuadas; e vogais em sílabas acentuadas possuem sua intensidade mais homogeneamente distribuída ao longo do espectro de frequências (SLUIJTER; HEUVEN, 1996). A sílaba acentuada pode ser também denominada de sílaba tônica. Dentre as línguas acentuais, há aquelas em que 1) o acento possui posição fixa, como o finlandês, em que a sílaba acentuada é sempre a primeira, e o polonês, em que a sílaba acentuada é sempre a penúltima; e 2) em que o acento ocupa posições variáveis, como o latim, o inglês e o português (HYMAN, 2009). Algumas dessas línguas acentuais, como o português e o inglês, fazem uso do acento com função contrastiva chamado, de modo análogo ao tom lexical, de acento lexical. Exemplos que ilustram o funcionamento do acento lexical podem ser os seguintes: 1) em português, as palavras sábia, sabia e sabiá possuem três significados diferentes, apesar de possuírem as mesmas sílabas, pois o que as diferencia é a posição do acento; 2) em inglês, os pares de palavras *below/billow* e *market/Marquette* também possuem significados diferentes de acordo com a posição do acento (LADEFOGED; JOHNSON, 2010, Cap. 10). As palavras exemplificadas acima, tanto em português quanto em inglês, possuem ortografias diferentes, mas são compostas pela mesma sequência de sons.

A oposição entre línguas tonais e acentuais não é tão clara, especialmente no caso de um grupo que línguas que empregam o *pitch accent*, termo em inglês que pode ser traduzido de forma livre para acento tonal. O *pitch accent* se manifesta por meio de variações de F0, assim como um tom lexical, mas sua função se assemelha à do acento, estabelecendo proeminência entre sílabas de uma mesma palavra. Línguas desse tipo, como o japonês e o servo-croata, empregam o *pitch accent* apenas em algumas palavras e com poucos padrões tonais (geralmente um ou dois), fazendo com que o uso do tom seja menos denso do que nas línguas tonais (YIP, M., 2002, Seção 9.3).

Deste modo, não se pode estabelecer uma divisão absoluta entre línguas tonais e acentuais. É mais coerente a ideia de que existe um espaço contínuo entre elas em que o emprego dos tons se faz cada vez menos denso e livre à medida que uma língua se torna mais acentual do que tonal ou vice-versa (YIP, M., 2002, Seção 1.1). A caracterização das línguas de *pitch accent* é, portanto, um tema de debate. Apesar da função do *pitch accent*

ser semelhante à do acento, essas línguas são geralmente classificadas como línguas tonais de acordo com a seguinte definição: 'Uma língua tonal é aquela em que uma indicação de F0 faz parte da realização lexical de pelo menos alguns morfemas' (HYMAN, 2001).

3.2 O tom lexical

Como início à discussão sobre tom lexical, é importante diferenciar de forma clara os seguintes termos: F0, altura e tom. Como detalhado no Capítulo 2, o termo F0, chamado de frequência fundamental, é controlado pela frequência de vibração das pregas vocais, que produzem uma onda sonora complexa composta por uma série harmônica de frequências múltiplas de F0, que é a frequência mais baixa dessa série. Por sua vez, o *pitch* (denominado em português como altura) é um termo psicofísico relacionado à percepção de F0: um sinal com valor alto de F0 é percebido como alto (ou agudo) enquanto que um sinal com valor baixo de F0 é percebido como baixo (ou grave). Contudo, essa relação entre F0 e *pitch* não é linear, pois os processamentos auditivos e cognitivos da fala não são lineares: uma variação de F0 não necessariamente muda, na mesma proporção, a percepção do *pitch* do som. O tom é 'um termo linguístico que se refere a uma categoria fonológica que distingue duas palavras ou sentenças' (YIP, M., 2002, p. 5). Como o termo tom pode ter diferentes significados em diferentes contextos, como por exemplo no da teoria musical, ao longo deste trabalho o tom será interpretado num contexto fonológico, para manter uniformidade de conceitos. Sumarizando sua relação com F0 e *pitch*, o tom se manifesta fisicamente por meio de F0 e é percebido por meio do *pitch*. Os três termos são, portanto, relevantes para o entendimento das línguas tonais e do tom lexical.

O parágrafo anterior apresentou neste trabalho o conceito de altura (em inglês *pitch*), diferenciando-o de F0. Até aqui, o tom foi descrito como o uso de padrões de variação de F0, mas é também possível descrevê-lo como o uso de padrões de variação de altura, pois F0 e altura são relacionadas. Para que o tom seja percebido, deve haver uma variação de F0 suficiente para que seja também percebida uma variação de altura. O modo de produção básico das variações de F0 foi descrito na Seção 2.1.3. Na Seção seguinte serão descritos outros aspectos que influenciam a produção dos tons, que podem ser intencionais ou não (OHALA, 1978).

3.2.1 Fatores de performance que afetam o tom

Os ciclos de vibração das pregas vocais que compõem a fonação ocorrem devido à existência de uma diferença suficiente entre as pressões subglótica e supraglótica. Na produção de consoantes oclusivas, quando há a obstrução completa das cavidades oral e nasal, essa diferença de pressão é pequena e pode não ser suficiente para que a fonação ocorra espontaneamente. Por outro lado, na produção de outros sons em que as cavidades

oral e/ou nasal estão abertas, a diferença de pressão é mais elevada, facilitando a fonação (YIP, M., 2002, Seção 1.2.1).

Devido à particularidade da produção das consoantes oclusivas, há um efeito do vozeamento delas no nível de F0 (e na altura) das vogais que as sucedem. Em línguas tonais, por exemplo, em que cada sílaba é pronunciada com um tom específico, percebe-se que o tom das vogais precedidas por consoantes oclusivas vozeadas possui nível mais baixo do que o tom das vogais precedidas por consoantes oclusivas não-vozeadas (OHALA, 1978; YIP, M., 2002, Seção 1.2.1). Fisiologicamente, acredita-se que a motivação desse fenômeno seja que consoantes oclusivas não-vozeadas são produzidas com as pregas vocais mais tensas, enquanto que as vozeadas são produzidas com as pregas vocais mais relaxadas (YIP, M., 2002, Seção 1.2.1).

Além disso, há também relação entre o parâmetro articulatório da altura da mandíbula na produção de vogais e o tom com que elas são pronunciadas. Sugere-se que essa relação seja a seguinte: quanto mais alta a mandíbula na articulação da vogal, maior seu tom. Acredita-se que a motivação fisiológica para isso seja que a articulação da língua em tais vogais tensiona as pregas vocais, aumentando o F0 do som produzido (OHALA, 1978).

Além dos dois fenômenos mencionados acima, serão expostos mais dois fenômenos da produção da fala que afetam os tons: o atraso de pico e a declinação. O atraso de pico acontece quando o pico de altura de um tom alto é atingido somente na sílaba seguinte à qual ele está linguisticamente acoplado. Isso se dá devido ao tempo de processamento, que se inicia com o envio de impulsos nervosos de cérebro até os músculos responsáveis pela produção do tom, que possui duração pequena, mas finita e relevante neste caso (YIP, M., 2002, Seção 1.2.2). Por outro lado, o fenômeno da declinação ocorre quando há um decaimento global do tom de uma sentença ao longo dela e é observado tanto em línguas tonais quanto em línguas não-tonais. Não há concordância global em relação ao mecanismo de produção do fenômeno de declinação, mas uma possível causa é a diminuição da massa de ar acumulada nos pulmões ao longo da sentença, que diminui a pressão subglótica e, consequentemente, F0 (OHALA, 1978; YIP, M., 2002, Seção 1.2.2).

A exposição dos fenômenos acima é uma evidência de que a produção do tom é um fenômeno complexo afetado por mais variáveis além daquelas relacionadas a seu mecanismo básico de produção, exposto na seção anterior, e que envolve questões que extrapolam os limites do modelo Fonte-Filtro, como será discutido posteriormente. Uma variação tonal pode, portanto, ter uma série de possíveis origens, intencionais ou não, e pode também ser percebida ou não como um padrão válido de variação. A seguir será detalhada a fonologia do tom, que descreve como o tom atua nas estruturas das línguas no mundo.

3.3 A fonologia do tom

Ao iniciar esta seção, convém conceituar fonética e fonologia. As duas áreas são comumente mencionadas juntas, mas procuram responder perguntas diferentes se utilizando de metodologias diferentes. A fonética se preocupa em descrever a estrutura física da fala por meio da descoberta e descrição dos sons vocais humanos, estudando sua articulação, sua acústica e sua percepção. Por outro lado, a fonologia se preocupa com os padrões e funções dos sons usados na linguagem e com a representação deles no cérebro (OHALA, 2010). Enquanto que as representações fonológicas são categóricas, sendo representadas de forma binária ou unária, as representações fonéticas são contínuas, em forma de gradientes. Um exemplo aplicado aos tons lexicais pode ser o seguinte: fonologicamente, a sílaba [ta] é especificada como sendo portadora de um tom alto (H). A fonética deve então interpretar o tom H levando, por exemplo, em consideração 1) a faixa de F0 na qual o falante em questão consegue falar; 2) em qual ponto dessa faixa o tom H será produzido (como o tom é alto, ele deve ser produzido nos valores mais altos da faixa); e 3) onde na fala o tom será produzido (como ele não pode ser produzido na consoante não-vozeada, ele será produzido na vogal) (YIP, M., 2002, Seção 1.3). Dessa forma, um mesmo tom alto falado por um homem em 170Hz e por uma mulher em 370Hz é fonologicamente igual, mas foneticamente diferente.

Até agora, este trabalho se preocupou majoritariamente com a fonética da fala e dos tons, descrevendo o modo como são produzidos e transmitidos. A seção seguinte, contudo, tratará da fonologia dos tons lexicais, explorando os modos por meio dos quais eles são representados e organizados nas línguas.

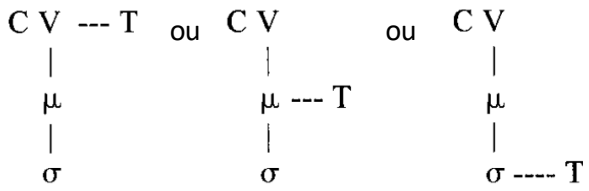
3.3.1 O tom descrito pela fonologia Autossegmental

A fonologia Autossegmental (GOLDSMITH, 1976) é uma teoria fonológica que representa os tons numa camada separada daquela dos segmentos. De acordo com sua representação autossegmental, o tom é um elemento autônomo associado a alguma unidade na camada segmental chamada de Unidade Portadora de Tom (do inglês *Tone Bearing Unit*, TBU). A TBU pode ser uma sílaba, uma vogal, uma mora e até mesmo uma consoante, no caso de algumas línguas. A Figura 10 ilustra três casos, cada um com uma TBU diferente: uma vogal, uma mora e uma sílaba, respectivamente. O tom pode se mover independentemente da camada segmental, pois ele existe em sua própria camada.

3.3.2 Notações tonais

Para publicar trabalhos sobre tom, são necessárias representações visuais, que no caso da fonética podem ser gráficos da variação de F0 ao longo do tempo e no caso da

Figura 10 – Possíveis TBUs de acordo com a fonologia Autossegmental. CV é uma sílaba consoante-vogal, sendo V a vogal, μ é uma mora e σ é uma sílaba. Nos três casos, a camada segmental está à esquerda e a camada dos tons à direita.



Fonte: Adaptado de (YIP, M., 2002, Cap. 4).

fonologia uma série de transcrições que podem variar entre si de acordo com a região geográfica de estudo (YIP, M., 2002, Seção 2.2).

No estudo de línguas africanas, a notação dos tons se dá por meio de acentos desenhados acima do segmento portador do tom, de acordo com a Tabela 1. Cada língua tonal possui um número específico de tons e utiliza as representações necessárias para representá-los. Uma língua com apenas dois tons, um alto e um baixo, precisa de apenas duas notações, enquanto que línguas mais complexas precisam de mais. Além dos tons descritos na Tabela 1, há ainda os tons extra-altos e extra-baixos, presentes em algumas línguas, que são representados por dois acentos agudos ou graves, respectivamente.

Tabela 1 – Representação fonológica dos tons de acordo com a tradição africana.

Tom	**Representação**
Alto	$\acute{a}$
Baixo	$\grave{a}$
Médio	$\bar{a}$
Descendente	$\hat{a}$
Crescente	$\check{a}$

Fonte: (YIP, M., 2002, Seção 2.2.1)

No estudo de línguas asiáticas, a notação dos tons não se dá por meio de acentos, mas por meio de números. O alcance de F0 natural de cada falante é dividido em 5 níveis, sendo o menor representado por 1 e o maior por 5. A cada sílaba são alocados de 0 a 3 dígitos: nenhum dígito caso ela não porte nenhum tom, 1 dígito para tons de nível de curta duração, 2 dígitos para tons de nível (dois números iguais) ou tons de contorno simples (o primeiro número indicando o tom inicial e o segundo o tom final) e 3 dígitos para tons de nível complexos, que possuem mudança de direção ao longo de sua duração (CHAO, 1930). A Tabela 2 traz alguns exemplos da notação de tons segundo a tradição asiática.

De modo semelhante ao estudo das línguas asiáticas, o estudos das línguas americanas também utiliza números de 1 a 5 para a notação de seus tons. A diferença, contudo, é

Tabela 2 – Representação fonológica dos tons de acordo com a tradição asiática.

Tom	**Representação**
Alto	ta^{55} ou ta^{5}
Médio	ta^{33} ou ta^{3}
Alto-crescente	ta^{35}
Baixo-descendente	ta^{31}
Baixo-descendente-crescente	ta^{214}
Baixo-crescente-descendente	ta^{231}

Fonte: (YIP, M., 2002, Seção 2.2.2)

que aqui o nível mais alto é representado pelo número 1 e o nível mais baixo pelo número 5 (YIP, M., 2002, Seção 2.2.3).

A seguir serão detalhadas as estruturas fonológicas dos tons lexicais nas três línguas utilizadas neste trabalho: cantonês, mandarim e tailandês. A região do sudeste asiático e do Pacífico é rica em línguas tonais, incluindo as famílias de línguas chinesa, papuana, tai-kadai e vietnamita, mas abriga também línguas não-tonais, como as faladas na Índia, Indonésia e Malásia (YIP, M., 2002, Cap. 7). O cantonês e o mandarim são línguas que fazem parte da família de línguas chinesa, juntamente com outras línguas como o taiwanês e o xangainês. Apesar de serem da mesma família, essas línguas não são mutuamente inteligíveis e possuem diferentes fonologias, léxicos e sintaxes. O tailandês, por sua vez, faz parte da família de línguas tai-kadai, juntamente com o laociano (língua falada em Laos) (YIP, M., 2002, Cap. 7).

Em comparação com as línguas tonais africanas, as línguas tonais do sudeste asiático e do Pacífico possuem um inventário tonal mais rico, composto não só por tons de nível, mas também por tons de contorno. Além disso, essas línguas possuem estruturas silábicas e morfológicas simples, e são os contrastes tonais que aumentam o inventário silábico significativamente, pois a mesma sílaba pronunciada com tons diferentes possui significados diferentes. Em mandarim, por exemplo, existem 406 sílabas segmentalmente diferentes, mas esse número chega a 1256 sílabas diferentes quando os contrastes tonais são levados em consideração (YIP, M., 2002, p. 172).

Para a descrição fonológica dos tons nas línguas estudadas neste trabalho, serão utilizadas duas notações, para manter o padrão utilizado nas descrições em (YIP, M., 2002): para as línguas chinesas cantonês e mandarim será utilizada a notação característica da tradição asiática, com números; e para o tailandês será utilizada uma notação semelhante, que indica tons altos pela letra H, tons médios pela letra M e tons baixos pela letra L.

3.3.3 A estrutura fonológica dos tons lexicais em cantonês

O cantonês padrão é falado em Hong Kong e no Cantão (*Guangzhou*) e é uma língua que faz parte da família de línguas chinesa, sendo, mais especificamente, um dialeto

yue do chinês (YIP, M., 2002, Seção 7.1). Possui sete tons lexicais diferentes nas sílabas com final sonoro e mais três nas sílabas com final oclusivo, como ilustrado na Tabela 3, que arbitrariamente numera os tons de 1 até 10 e os nomeia de acordo com seu contorno ou nível.

Nas sílabas com final sonoro, os tons 1 e 7 não são distintos para a maioria dos falantes, resultando, portanto, em seis tons lexicais diferentes. Nas sílabas com final oclusivo ocorrem três tons de nível, correspondentes aos tons 8, 9 e 10, nos mesmo níveis que ocorrem nas sílabas com final sonoro, mas com duração menor: a notação desses tons é feita com apenas um número ao invés de dois. A TBU em cantonês é a mora, com tons de nível sendo portados por uma mora e tons de contorno (tons com dois níveis distintos) sendo portados por duas moras (YIP, M., 2002, Seção 7.1).

Tabela 3 – Estrutura fonológica dos tons do cantonês padrão.

Tom	**Denominação**	**Notação**
1	Alto	55
2	Crescente	35
3	Médio	44
4	Baixo-descendente	22/21
5	Baixo-crescente	24
6	Baixo	33
7	Alto-descendente	53
8	Alto curto	5
9	Médio curto	4
10	Baixo curto	3

Fonte: (YIP, M., 2002, Seção 7.1)

3.3.4 A estrutura fonológica dos tons lexicais em mandarim

A maior família dentre as línguas chinesas é a do mandarim, que inclui os dialetos *tianjin* e mandarim de *Beijing* (YIP, M., 2002, Seção 7.2). Esta seção abordará o mandarim de *Beijing*, nomeando-o apenas de mandarim, por simplificação. As línguas da família mandarim possuem, no geral, menos tons do que as línguas da família yue: o mandarim, por exemplo, possui quatro tons distintos, detalhados na Tabela 4.

Há uma particularidade relativa ao tom 3 chamada de sândi tonal, que é o efeito dos tons adjacentes sobre um determinado tom. No caso do mandarim, o tom 3 ocorre como 21 em posições não-finais da sentença, e ocorre como 214 em posições de fim de sentença (YIP, M., 2002, Seção 7.2). Além disso, no mandarim há também uma discussão sobre o tom neutro, que ocorre em sílabas curtas que não portam nenhum tom. Há casos em que tais sílabas nunca portam tom e casos em que elas podem portar tom a depender do contexto. Além disso, a TBU em mandarim é, assim como no cantonês, a mora: cada mora pode portar um tom (YIP, M., 2002, Seção 7.2).

Tabela 4 – Estrutura fonológica dos tons do mandarim.

Tom	**Denominação**	**Notação**
1	Alto	55
2	Crescente	35
3	Baixo-descendente-crescente	21(4)
4	Descendente	53

Fonte: (YIP, M., 2002, Seção 7.2)

3.3.5 A estrutura fonológica dos tons lexicais em tailandês

O tailandês é uma língua que faz parte da família de línguas tai-kadai e que possui 5 tons lexicais diferentes, de acordo com a Tabela 5. De modo semelhante ao cantonês, no tailandês sílabas com final oclusivo somente podem portar alguns tons: se forem com vogais longas, somente os tons 2 e 3; se foram com vogais curtas, somente os tons 2 e 4. De forma similar ao cantonês e ao mandarim, a TBU em tailandês é a mora (YIP, M., 2002, Seção 7.7).

Tabela 5 – Estrutura fonológica dos tons do tailandês.

Tom	**Denominação**	**Notação**
1	Descendente	HL
2	Baixo	L
3	Alto	H
4	Médio	M
5	Crescente	LH

Fonte: (YIP, M., 2002, Seção 7.7)

3.4 O caráter multimodal do tom lexical

Esta seção tem como objetivo contextualizar os conceitos apresentados até aqui, nos capítulos 2 e 3, com estudos de fala multimodal em línguas tonais, evidenciando a importância da componente visual da fala para a produção e para a percepção de tons lexicais.

Conforme exposto na Seção 2.1.3, há mecanismos na laringe que controlam o nível de F0 produzido. Por exemplo, a contração do músculo CT é relacionada ao aumento de F0, enquanto que o abaixamento de F0 pode estar relacionado tanto ao relaxamento do músculo CT aliado à contração do músculo TA, quanto ao abaixamento da global da laringe, que envolve músculos extra-laríngeos como o tireo-hioideo (TH), o esterno-hioideo (SH) e o esterno-tireoideo (ST), ilustrados na Figura 7.

De acordo com o modelo Fonte-Filtro (FANT, 1960), apresentado na Seção 2.2.1, a fonte sonora, representada pela fonação, é independente do filtro, que corresponde ao trato

vocal e que se reflete, por exemplo, nas frequências formantes do sinal de fala resultante. Contudo, faz sentido observar que essa independência se encontra no plano ideal e não no real. A mudança na altura global da laringe, que é um fator que controla F0, tem como consequência uma mudança no comprimento do trato vocal: se a laringe é rebaixada, o comprimento do trato vocal aumenta e vice-versa. Essa relação, por exemplo, ilustra uma mudança no filtro causada pela fonte.

Há estudos que investigam com maior profundidade essa dependência e que cunharam o termo F0 intrínseco (I_{F0}), que é a tendência de vogais altas terem F0maior do que vogais baixas. Esse fenômeno é universal e foi encontrado em todas as línguas que foram objetos de estudo até então, tonais ou não-tonais (WHALEN; LEVITT, 1995). Esse fenômeno, já também descrito na Seção 3.2.1 (OHALA, 1978; YIP, M., 2002, Seção 2.6), sistematiza a interação entre a produção de tons e vogais, correlatos da fonte e do filtro respectivamente.

Essa relação é de particular interesse para as línguas tonais, pois nelas o F0 não pode variar livremente, já que possui efeito contrastivo entre palavras e deve obedecer a padrões fonológicos. Dentre as perguntas norteadoras dos estudos de I_{F0} em línguas tonais e de suas ramificações, podem ser destacadas as seguintes: 1) se a posição de articuladores como a língua, os lábios e a mandíbula muda sistematicamente com o tom, para uma mesma vogal (ERICKSON et al., 2014; HU, 2004); 2) se a percepção dos tons lexicais está conectada de alguma forma à informação visual do falante (BURNHAM; CIOCCA; STOKES, 2001; BURNHAM; LAU et al., 2001); e 3) como se dá o fenômeno de I_{F0} em línguas tonais (ZEE, 1980).

Dentre os resultados obtidos sobre a relação entre a produção de vogais e os tons lexicais, podem ser destacados os seguintes: em taiwanês, a articulação das vogais é afetada pelos diferentes tons lexicais, mas não de modo sistemático comum a todos os falantes (ZEE, 1980). Em dados de mandarim e de ningbo, outra língua chinesa, foi encontrada relação entre tom lexical, posição do maxilar e da língua e a primeira frequência formante (F1): na produção do tom 3, o mais baixo, o maxilar e a língua se encontraram mais retraídos e o valor de F1 foi maior do que na produção do tom 1, o mais alto (ERICKSON et al., 2014; HU, 2004). Em dados de mandarim foram observadas diferenças articulatórias e de movimentação da cabeça entre vogais produzidas com o tom 3 e vogais produzidas com os demais tons: retração e abaixamento da língua e da mandíbula e uma tendência de maior velocidade de movimentos da cabeça para frente e para cima diferenciaram o tom 3 dos demais (HOOLE; HU, 2004). Os resultados acima, contudo, não foram generalizados para todas as vogais e foram obtidos com um número baixo de falantes. Eles, portanto, apenas indicam que possa haver relações sistemáticas entre produção de tons lexicais e a articulação de vogais e o movimento da cabeça, pois ainda não foram realizados estudos exaustivos sobre o tema.

Pesquisas mais recentes obtiveram resultados investigando a percepção dos tons lexicais em função da vogal portadora, observando que a vogal condiciona a rapidez de reconhecimento do tom: um tom crescente, por exemplo, é reconhecido mais rapidamente quando falado com a vogal [i] do que quando falado com a vogal [u] (SHAW, Jason A et al., 2013). A partir desses resultados surge a especulação de que movimentos específicos na articulação das vogais poderiam providenciar informações sobre o tom lexical de forma mais rápida do que o próprio F0, já que tons e vogais são sobrepostos temporalmente e a articulação da vogal se dá de forma mais rápida do que a produção do tom (SHAW, Jason A. et al., 2016). Dados esses resultados, (SHAW, Jason A. et al., 2016) ainda sugerem uma abordagem mais apropriada para a integração entre vogais e tons lexicais: cada combinação de tom e vogal deve possuir uma articulação única, que pode resultar em benefícios perceptivos.

Há, portanto, evidências que justificam este trabalho, que procura estudar correlatos visuais da produção de tons lexicais. Partindo-se das evidências de que há relação entre a produção de tons e a articulação de vogais e de que a articulação de vogais influencia o movimento da face por meio de articuladores como mandíbula, lábios e língua, então conclui-se que pode haver relação entre a produção de tons e o movimento facial.

No nível da sentença, onde o tom se manifesta por meio da entoação, já foram observados correlatos visuais na face e na cabeça (BURNHAM; CIOCCA; STOKES, 2001). Em francês, uma língua não-tonal, foi encontrada correlação entre a entoação e o movimento das sobrancelhas dos falantes (CAVE et al., 1996). Além disso, fortes relações entre o movimento da cabeça e o nível de F0 foram também observadas (YEHIA; KURATATE; VATIKIOTIS-BATESON, 2002), com relações mais específicas com unidades prosódicas mais longas do que o tom lexical, como o *pitch accent* (KRAHMER; SWERTS, 2007). Os resultados acima, relacionando outro fenômeno baseado no tom com movimentos faciais e de cabeça, também justificam o objetivo deste trabalho, que busca estudar a mesma relação em tons lexicais.

4 Base de dados

Este capítulo apresenta a base de dados utilizada neste trabalho. Foram adquiridos dados audiovisuais em experimentos de produção da fala conduzidos com falantes de três línguas tonais: cantonês, mandarim e tailandês. São descritos os experimentos para aquisição dos dados, assim como o processamento realizado nos dados para que ficassem no formato necessário aos algoritmos de classificação descritos no Capítulo 5 e utilizados no Capítulo 6.

4.1 Aquisição dos dados

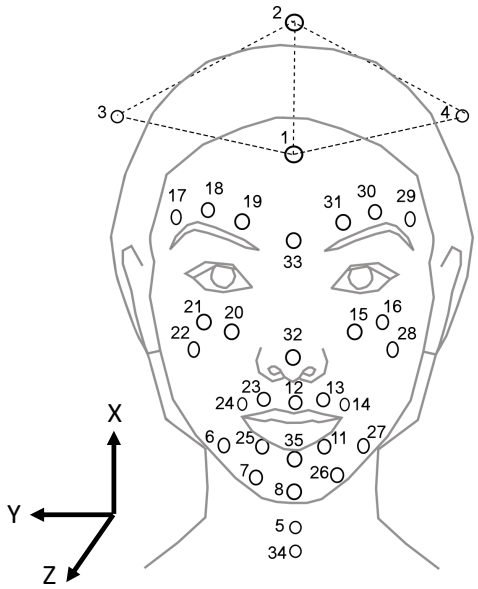
Os experimentos realizados para a aquisição dos dados foram realizados pela equipe do professor Denis Burnham, do Instituto MARCS de pesquisa em cérebro, comportamento e desenvolvimento (*MARCS Institute for Brain, Behavior and Development*), da *Western Sydney University*, em Sydney, Austrália. Foram realizados experimentos individuais com sete participantes, sendo três falantes nativas de cantonês, três falantes nativas de mandarim e uma falante nativa de tailandês, todas do gênero feminino e com idade entre 20 e 30 anos.

Por se tratar de um experimento audiovisual, foram realizados dois tipos de medições: a acústica da fala, usando um microfone, e o movimento da face/cabeça, usando marcadores colocados na face das participantes que foram rastreados pelo OPTOTRAK (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020a). Conforme exposto no Capítulo 1, o OPTOTRAK é um equipamento que rastreia a posição de marcadores emissores de luz infravermelha, também chamados de marcadores ativos. A aquisição dos sinais de movimento foi realizada pelo OPTOTRAK a uma frequência de amostragem de 60Hz. Cada amostra registra a posição de cada marcador em três coordenadas (x, y, z). Sendo no total 33 marcadores, de acordo com a Figura 11, cada amostra coleta $33 \times 3 = 99$ valores de posição, em milímetros. A aquisição dos sinais acústicos foi realizada simultaneamente à aquisição dos sinais de movimento por meio de um microfone de alta qualidade. A sincronização entre os sinais acústicos e de movimento foi realizada por meio da conexão da saída analógica do microfone ao equipamento ODAU (OPTOTRAK *Data Acquisition Unit*) (NORTHERN DIGITAL MEASUREMENT SCIENCES, 2020b), que sincroniza o áudio com as câmeras do OPTOTRAK e o digitaliza numa frequência de amostragem de 44100Hz.

Durante os experimentos, cada participante permaneceu sentada numa cadeira a cerca de $2,0$m de distância do OPTOTRAK com 33 marcadores fixados em sua cabeça, face e pescoço, de acordo com a Figura 11. Após ouvir um bipe, produzido pela condutora

do experimento, cada participante falava a palavra que aparecia numa tela posicionada diretamente à sua frente.

Figura 11 – Posições dos marcadores do OPTOTRAK na face das participantes. Os marcadores 1 a 4 foram presos a um capacete utilizado pelas participantes e são designados para capturar o movimento de corpo rígido da cabeça. Os marcadores 5 e 34 foram colocados no pescoço das participantes. Os marcadores 9 e 10 estavam inativos durante os experimentos.



Fonte: Cantoni, Maria Mendes

Para cada língua, o *corpus* coletado foi diferente, mas tendo sempre em vista uma variedade linguística abrangente, com diversos tipos de vogais combinados com diversos tipos de consoantes. Em **cantonês**, cada uma das três participantes produziu 216 palavras e sílabas diferentes, sendo:

- 12 palavras × 6 tons = 72 produções;
- 24 sílabas (8 consoantes combinadas com 3 vogais) × 6 tons = 144 produções

Cada participante repetiu cada produção 4 vezes, resultando em 864 produções por participante. Uma descrição das produções pode ser encontrada na Tabela 6. As três participantes falantes nativas de cantonês serão denominadas ao longo do trabalho como CNT1, CNT2 e CNT3.

Em **mandarim**, cada participante produziu 168 palavras e sílabas diferentes, sendo:

- 18 palavras × 4 tons = 72 produções;

Tabela 6 – Descrição das elocuções de cantonês. Uma maior descrição sobre os tipos de vogais e de consoantes pode ser vista no Capítulo 2.

Produção	**Tipo de vogal**	**Tipo de consoante**
[ji]	Alta anterior	Aproximante palatal vozeada
[fu]	Alta posterior	Fricativa labiodental não-vozeada
[si]	Alta anterior	Fricativa alveolar não-vozeada
[sɛ]	Média anterior	Fricativa alveolar não-vozeada
[fɐn]	Baixa central nasal	Fricativa labiodental não-vozeada
[jen]	Alta anterior nasal	Aproximante palatal vozeada
[hau]	Alta anterior	Fricativa glotal não-vozeada
[ha:u]	Ditongo	Fricativa glotal não-vozeada
[jau]	Ditongo	Aproximante palatal vozeada
[soei]	Ditongo	Fricativa alveolar não-vozeada
[wai]	Ditongo	Aproximante labiovelar vozeada
[p^ha]	Baixa central	Oclusiva bilabial aspirada não-vozeada
[p^hi]	Alta anterior	Oclusiva bilabial aspirada não-vozeada
[p^hu]	Alta posterior	Oclusiva bilabial aspirada não-vozeada
[pa]	Baixa central	Oclusiva bilabial não-vozeada
[pi]	Alta anterior	Oclusiva bilabial não-vozeada
[pu]	Alta posterior	Oclusiva bilabial não-vozeada
[k^ha]	Baixa central	Oclusiva velar aspirada não-vozeada
[k^hi]	Alta anterior	Oclusiva velar aspirada não-vozeada
[k^hu]	Alta posterior	Oclusiva velar aspirada não-vozeada
[ka]	Baixa central	Oclusiva velar não-vozeada
[ki]	Alta anterior	Oclusiva velar não-vozeada
[ku]	Alta posterior	Oclusiva velar não-vozeada
[t^ha]	Baixa central	Oclusiva alveolar aspirada não-vozeada
[t^hi]	Alta anterior	Oclusiva alveolar aspirada não-vozeada
[t^hu]	Alta posterior	Oclusiva alveolar aspirada não-vozeada
[ta]	Baixa central	Oclusiva alveolar não-vozeada
[ti]	Alta anterior	Oclusiva alveolar não-vozeada
[tu]	Alta posterior	Oclusiva alveolar não-vozeada
[ma]	Baixa central	Nasal bilabial vozeada
[mi]	Alta anterior	Nasal bilabial vozeada
[mu]	Alta posterior	Nasal bilabial vozeada
[na]	Baixa central	Nasal alveolar vozeada
[ni]	Alta anterior	Nasal alveolar vozeada
[nu]	Alta posterior	Nasal alveolar vozeada

- 24 sílabas (8 consoantes combinadas com 3 vogais) $\times$ 4 tons = 96 produções

Cada participante repetiu cada produção 5 vezes, resultando em 840 produções por participante. Uma descrição das produções pode ser encontrada na Tabela 7. As três participantes falantes nativas de mandarim serão denominadas ao longo do trabalho como MND1, MND2 e MND3.

Tabela 7 – Descrição das produções de mandarim. Uma maior descrição sobre os tipos de vogais e os modos de articulação de consoantes pode ser vista no Capítulo 2.

Produção	**Tipo de vogal**	**Modo de articulação da consoante**
[ji]	Alta anterior	Aproximante palatal vozeada
[fu]	Alta posterior	Fricativa labiodental não-vozeada
[fan]	Baixa central nasal	Fricativa labiodental não-vozeada
[hao]	Ditongo	Fricativa glotal não-vozeada
[wei]	Ditongo	Aproximante labiovelar vozeada
[pao]	Ditongo	Oclusiva bilabial não-vozeada
[ts^{h}ai]	Ditongo	Africada dento-alveolar não-vozeada
[fɤn]	Alta posterior	Fricativa labiodental não-vozeada
[jü]	Alta posterior	Aproximante palatal vozeada
[tɕü]	Alta posterior	Africada alveolo-palatal não-vozeada
[nü]	Alta posterior	Nasal alveolar vozeada
[tü]	Alta posterior	Oclusiva alveolar não-vozeada
[p^ho]	Média central	Oclusiva bilabial aspirada não-vozeada
[po]	Média central	Oclusiva bilabial não-vozeada
[mo]	Média central	Nasal bilabial vozeada
[p^ha]	Baixa central	Oclusiva bilabial aspirada não-vozeada
[p^hi]	Alta anterior	Oclusiva bilabial aspirada não-vozeada
[p^hu]	Alta posterior	Oclusiva bilabial aspirada não-vozeada
[pa]	Baixa central	Oclusiva bilabial não-vozeada
[pi]	Alta anterior	Oclusiva bilabial não-vozeada
[pu]	Alta posterior	Oclusiva bilabial não-vozeada
[k^ha]	Baixa central	Oclusiva velar aspirada não-vozeada
[k^hi]	Alta anterior	Oclusiva velar aspirada não-vozeada
[k^hu]	Alta posterior	Oclusiva velar aspirada não-vozeada
[ka]	Baixa central	Oclusiva velar não-vozeada
[ki]	Alta anterior	Oclusiva velar não-vozeada
[ku]	Alta posterior	Oclusiva velar não-vozeada
[t^ha]	Baixa central	Oclusiva alveolar aspirada não-vozeada
[t^hi]	Alta anterior	Oclusiva alveolar aspirada não-vozeada
[t^hu]	Alta posterior	Oclusiva alveolar aspirada não-vozeada
[ta]	Baixa central	Oclusiva alveolar não-vozeada
[ti]	Alta anterior	Oclusiva alveolar não-vozeada
[tu]	Alta posterior	Oclusiva alveolar não-vozeada
[ma]	Baixa central	Nasal bilabial vozeada
[mi]	Alta anterior	Nasal bilabial vozeada
[mu]	Alta posterior	Nasal bilabial vozeada
[na]	Baixa central	Nasal bilabial vozeada
[ni]	Alta anterior	Nasal bilabial vozeada
[nu]	Alta posterior	Nasal bilabial vozeada

Em **tailandês**, a participante produziu 280 palavras e sílabas diferentes, sendo:

- 12 palavras × 5 tons = 60 produções;
- 44 sílabas (11 consoantes combinadas com 3 vogais) × 5 tons = 220 produções

A participante repetiu cada produção 4 vezes, resultando em 1120 produções. Uma descrição das produções pode ser encontrada nas Tabelas 8 e 9. A participante falante nativa de tailandês será denominada ao longo do trabalho como TAI1.

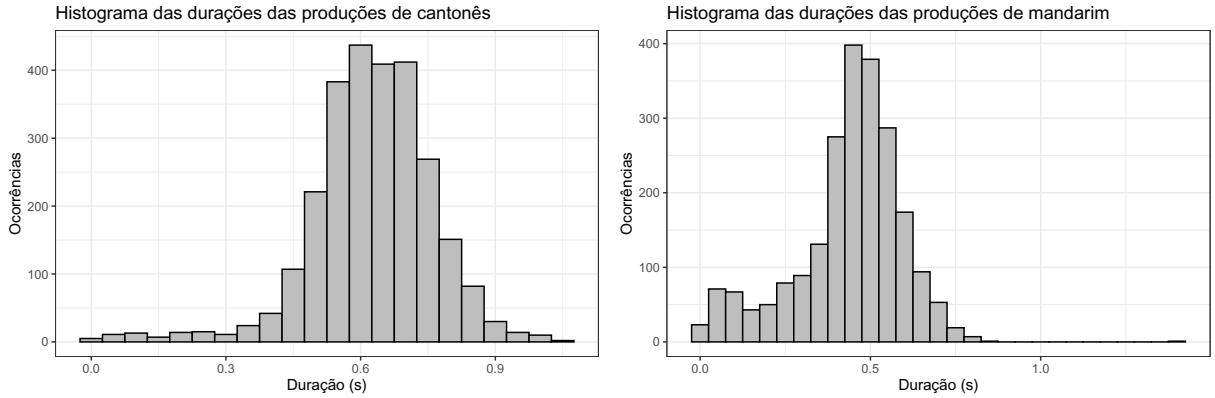
Tabela 8 – Descrição das produções de palavras em tailandês. Uma maior descrição sobre os tipos de vogais e os modos de articulação de consoantes pode ser vista no Capítulo 2.

Produção	**Tipo de vogal**	**Modo de articulação da consoante**
[ji]	Alta anterior	Aproximante palatal vozeada
[fu]	Alta posterior	Fricativa labiodental não-vozeada
[pau]	Ditongo	Oclusiva bilabial não-vozeada
[hau]	Ditongo	Fricativa glotal não-vozeada
[tɕ^{h}ai]	Ditongo	Africada palatal não-vozeada
[fan]	Baixa central nasal	Fricativa labiodental não-vozeada
[wai]	Ditongo	Aproximante velar vozeada
[si:]	Alta posterior	Fricativa alveolar não-vozeada
[sɛ:]	Baixa anterior	Fricativa alveolar não-vozeada
[jan]	Baixa central nasal	Aproximante palatal vozeada
[soej]	Ditongo	Fricativa alveolar não-vozeada

Tabela 9 – Descrição das produções de sílabas em tailandês. Uma maior descrição sobre os tipos de vogais e os modos de articulação de consoantes pode ser vista no Capítulo 2.

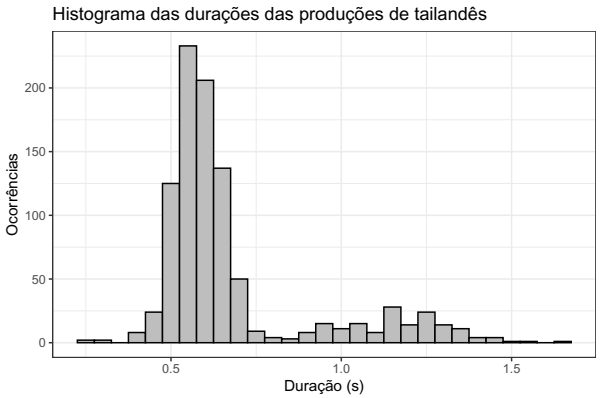
Produção	**Tipo de vogal**	**Modo de articulação da consoante**
[p^ha:]	Baixa central	Oclusiva bilabial aspirada não-vozeada
[p^hi:]	Alta anterior	Oclusiva bilabial aspirada não-vozeada
[p^hu:]	Alta posterior	Oclusiva bilabial aspirada não-vozeada
[p^hɯ:]	Alta central	Oclusiva bilabial aspirada não-vozeada
[pa:]	Baixa central	Oclusiva bilabial não-vozeada
[pi:]	Alta anterior	Oclusiva bilabial não-vozeada
[pu:]	Alta posterior	Oclusiva bilabial não-vozeada
[pɯ:]	Alta central	Oclusiva bilabial não-vozeada
[ba:]	Baixa central	Oclusiva bilabial vozeada
[bi:]	Alta anterior	Oclusiva bilabial vozeada
[bu:]	Alta posterior	Oclusiva bilabial vozeada
[bɯ:]	Alta central	Oclusiva bilabial vozeada
[t^ha:]	Baixa central	Oclusiva alveolar aspirada não-vozeada
[t^hi:]	Alta anterior	Oclusiva alveolar aspirada não-vozeada
[t^hu:]	Alta posterior	Oclusiva alveolar aspirada não-vozeada
[t^hɯ:]	Alta central	Oclusiva alveolar aspirada não-vozeada
[ta:]	Baixa central	Oclusiva alveolar não-vozeada
[ti:]	Alta anterior	Oclusiva alveolar não-vozeada
[tu:]	Alta posterior	Oclusiva alveolar não-vozeada
[tɯ:]	Alta central	Oclusiva alveolar não-vozeada
[da:]	Baixa central	Oclusiva alveolar vozeada
[di:]	Alta anterior	Oclusiva alveolar vozeada
[du:]	Alta posterior	Oclusiva alveolar vozeada
[dɯ:]	Alta central	Oclusiva alveolar vozeada
[k^ha:]	Baixa central	Oclusiva velar aspirada não-vozeada
[k^hi:]	Alta anterior	Oclusiva velar aspirada não-vozeada
[k^hu:]	Alta posterior	Oclusiva velar aspirada não-vozeada
[k^hɯ:]	Alta central	Oclusiva velar aspirada não-vozeada
[ka:]	Baixa central	Oclusiva velar não-vozeada
[ki:]	Alta anterior	Oclusiva velar não-vozeada
[ku:]	Alta posterior	Oclusiva velar não-vozeada
[kɯ:]	Alta central	Oclusiva velar não-vozeada
[ma:]	Baixa central	Nasal bilabial vozeada
[mi:]	Alta anterior	Nasal bilabial vozeada
[mu:]	Alta posterior	Nasal bilabial vozeada
[mɯ:]	Alta central	Nasal bilabial vozeada
[na:]	Baixa central	Nasal alveolar vozeada
[ni:]	Alta anterior	Nasal alveolar vozeada
[nu:]	Alta posterior	Nasal alveolar vozeada
[nɯ:]	Alta central	Nasal alveolar vozeada
[ŋa:]	Baixa central	Nasal velar vozeada
[ŋi:]	Alta anterior	Nasal velar vozeada
[ŋu:]	Alta posterior	Nasal velar vozeada
[ŋɯ:]	Alta central	Nasal velar vozeada

Figura 12 – Histograma das durações das produção em cantonês, mandarim e tailandês.



(a) Cantonês.

(b) Mandarim.



(c) Tailandês.

Fonte: o autor

As palavras e sílabas produzidas por cada participante do experimento possuem durações diferentes. Como os métodos de classificação estatística que serão aplicados necessitam que os dados tenham todos a mesma duração para serem organizados numa matriz (cada linha é uma observação com um mesmo número de colunas - variáveis - associado a cada uma), isto terá de ser resolvido. A Seção 4.3 descreve o procedimento realizado para isso. A Figura 12 e a Tabela 10 ilustram a distribuição das durações das produções para cada umas das três línguas.

Tabela 10 – Média, desvio padrão e valores máximo e mínimo de duração das produções em cantonês, mandarim e tailandês. Todos os valores estão descritos em segundos.

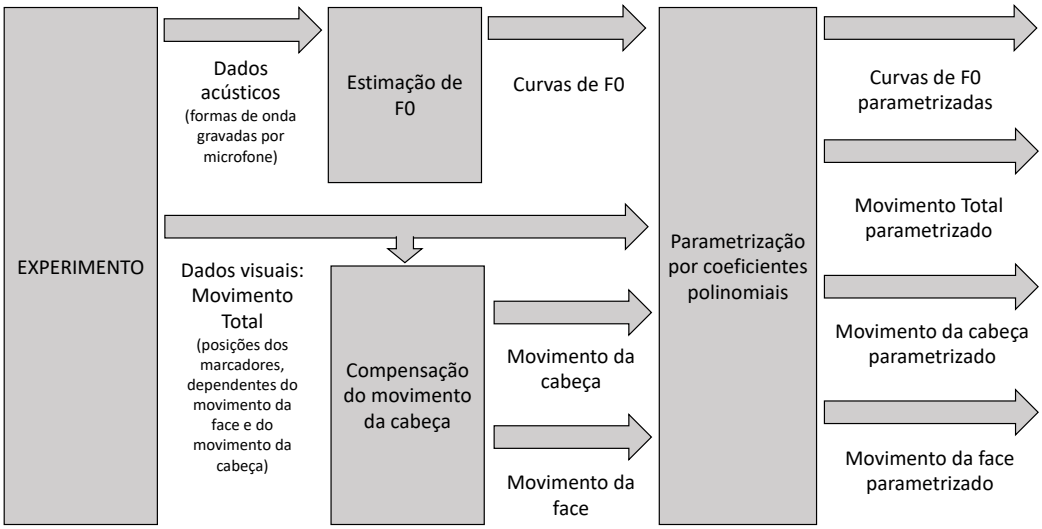
Língua	**Média**	**Desvio padrão**	**Valor máximo**	**Valor mínimo**
Cantonês	0,63	0,14	1,07	0,02
Mandarim	0,44	0,16	1,42	0,02
Tailandês	0,69	0,23	1,63	0,23

4.2 Processamento dos dados

Sinais podem ser de difícil processamento em sua forma crua, isto é, logo após sua aquisição. Um exemplo disso é um sinal de fala, que logo após ser gravado é descrito pela intensidade de suas oscilações no domínio do tempo. Muitas aplicações práticas não utilizam os sinais crus, mas os processam de modo que propriedades específicas sejam ressaltadas, facilitando sua interpretação. É nesse sentido que as técnicas de extração de características se fazem importantes: extraindo características do sinal relevantes para cada aplicação, como por exemplo sua intensidade média, sua variância ou sua composição espectral.

A base de dados descrita na seção anterior é formada por sinais acústicos e por sinais de movimento, e cada um desses tipos de sinal passa por processamentos diferentes. Os sinais acústicos serão representados por F0, que é o correlato acústico mais relacionado a tons. Os sinais de movimento passarão pelo procedimento de compensação do movimento da cabeça, que separa as componentes de movimento da face e da cabeça. Por fim, ambos os sinais terão suas durações normalizadas ao longo de todas as elocuções, o que é uma exigência dos métodos de classificação a serem aplicados. A Figura 13 ilustra os procedimentos realizados nos dados que serão descritos no restante deste capítulo.

Figura 13 – Diagrama de blocos representativo do processamento da base de dados. O experimento foi descrito na Seção 4.1. A estimação de F0 será descrita na Seção 4.2.1 e a compensação do movimento da cabeça será descrita na Seção 4.2.2. A parametrização dos sinais será descrita na Seção 4.3.



Fonte: o autor

4.2.1 Processamento dos sinais acústicos

Os tons lexicais são geralmente caracterizados pela frequência fundamental (F0) (YIP, M., 2002). Portanto, ao invés de usarmos a forma de onda dos sinais acústicos para a classificação de tons lexicais, estimaremos a curva de F0 a partir dessas formas de onda e utilizaremos os sinais de F0 como sinal de entrada para os classificadores. Para isso, estimaremos os valores de F0 correspondentes a cada sinal acústico da base de dados.

Como visto no Capítulo 2, as ondas sonoras da fala são compostas por trechos de relativa periodicidade (trechos quasi-periódicos tratados na prática como periódicos) e por trechos de aperiodicidade. Nos trechos tratados como periódicos há sobreposição de componentes periódicas de diversas frequências, todas múltiplas de uma frequência fundamental F0, ao passo que nos trechos aperiódicos isso não ocorre. Deste modo, F0 existe e pode ser estimada somente nos trechos tratados como periódicos.

A estimação de F0 é às vezes chamada de detecção de F0, apesar de estimação e detecção serem problemas diferentes: um problema de detecção é sobre aceitar ou rejeitar uma hipótese ou série de hipóteses enquanto que um problema de estimação se refere a determinar um certo número ou quantidade. Ao lidar com F0, existem os dois problemas: estimação e detecção: determinar o valor específico de F0 num determinado momento é um problema de estimação, enquanto que determinar se F0 existe ou não (se a fala é vozeada ou desvozeada, por exemplo) é um problema de detecção (CHRISTENSEN; JAKOBSSON, 2009, Seção 1.1). Aqui trataremos do problema de estimação de F0.

Existem diversos métodos para estimação de F0(CHRISTENSEN; JAKOBSSON, 2009), mas focaremos apenas no método da autocorrelação, especificamente naquele descrito por (BOERSMA, 1993). Esse método se baseia na F0 de um sinal em dado momento pode ser encontrado pela posição do valor máximo da função de autocorrelação do sinal.

Um fator importante para a definição e para a implementação do método é que o sinal de fala não é estacionário[1], pois sua distribuição estatística depende principalmente do som que está sendo produzido, variando a cada fonema. Um sinal de fala de uma pessoa falando a vogal [a] de forma prolongada e constante é um sinal estacionário, mas que não é usualmente utilizado em situações reais de comunicação. Um sinal de fala mais comum, de uma pessoa falando uma palavra, por exemplo, já não é estacionário, pois sua distribuição é modificada a cada som. Lançando mão do modelo Fonte-filtro (FANT, 1960) e de que a fala é um fenômeno em que tanto a fonte quanto o filtro se modificam continuamente, pode-se dizer mudanças na configuração da fonte (vozeada em vogais e não vozeada em consoantes) e do filtro (formatos diferentes para sons diferentes) modificam a distribuição estatística do sinal de fala também continuamente.

[1] Um sinal é estacionário quando sua distribuição estatística (determinada pelas propriedades estatísticas, como valor esperado e variância) é constante ao longo de sua duração (PAPOULIS, 2002, Cap. 9)

Para que a estimação de F0 baseada no método da autocorrelação (BOERSMA, 1993) seja aplicável a sinais de fala, é utilizada a operação de autocorrelação de tempo curto, aplicada em quadros de pequena duração, geralmente sobrepostos entre si, do sinal de fala completo $x(t)$. Pressupondo que as propriedades estatísticas do sinal de fala variam de forma lenta, é possível considerar cada um desses quadros de curta duração como um sinal estacionário (RABINER; SCHAFER, 2011, Seção 6.2). Deste modo, um valor de F0 é estimado para cada quadro de $x(t)$. Esta divisão do sinal de fala em quadros é feita em vários cenários como, por exemplo, para a estimação dos parâmetros do modelo Fonte-filtro (parâmetros LPC (RABINER; SCHAFER, 2011, Cap. 9)).

Para implementar o método, é realizada primeiramente a extração de um pequeno segmento de $x(t)$, denominado de quadro, com duração T (a duração da janela, $24ms$ na Figura 14) centralizado em t_{medio} ($12ms$ na Figura 14). A média do quadro μ_x é subtraída do quadro, que é então multiplicado por uma janela $w(t)$, resultando no sinal janelado

$$a(t) = \left(x\left(t_{medio} - \frac{T}{2} + t\right) - \mu_x\right) w(t). \tag{4.1}$$

A janela utilizada no método descrito é a janela de Hanning, simétrica em torno de $t = \frac{T}{2}$ e nula fora do intervalo $[0, T]$, dada por

$$w(t) = \frac{1}{2} - \frac{1}{2}\cos\left(\frac{2\pi t}{T}\right). \tag{4.2}$$

A multiplicação do quadro de $x(t)$ pela janela $w(t)$ resulta na suavização do sinal janelado $a(t)$ em suas extremidades, minimizando problemas de continuidade entre quadros adjacentes. Após a obtenção de $a(t)$, é calculada sua autocorrelação normalizada $r_a(\tau)$:

$$r_a(\tau) = \frac{\int_0^T a(t)a(t+\tau)dt}{\int_0^T a(t)a(t)dt}, \tag{4.3}$$

sendo τ correspondente a valores de atraso do sinal janelado $a(t)$. A obtenção da autocorrelação normalizada $r_x(\tau)$ de cada quadro de $x(t)$ é então obtida da seguinte forma:

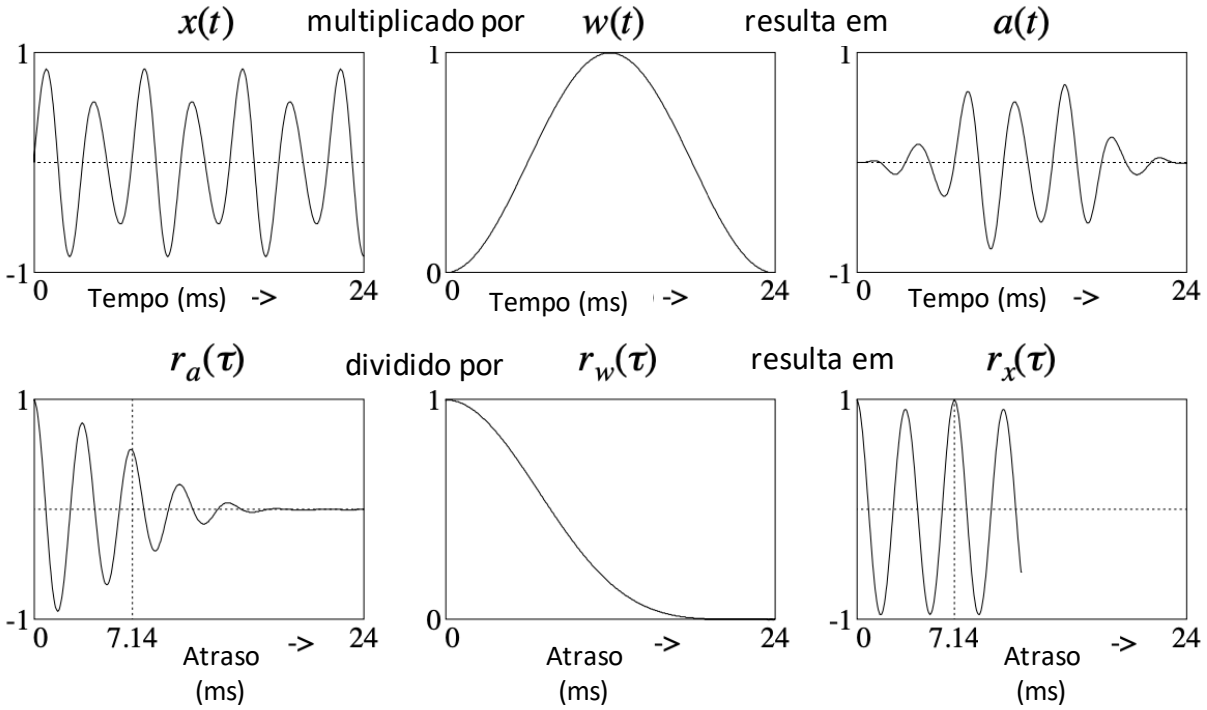
$$r_x(\tau) = \frac{r_a(\tau)}{r_w(\tau)}, \tag{4.4}$$

sendo $r_w(\tau)$ a autocorrelação normalizada da janela de Hanning $w(t)$ descrita por

$$r_w(\tau) = \left(1 - \frac{|\tau|}{T}\right)\left(\frac{2}{3} + \frac{1}{3}\cos\frac{2\pi\tau}{T}\right) + \frac{1}{2\pi}\sin\frac{2\pi|\tau|}{T}. \tag{4.5}$$

As operações descrita pelas equações 4.1 e 4.4 são ilustradas na Figura 14.

Figura 14 – Nos três gráficos superiores: janelamento do sinal $x(t)$ pela janela $w(t)$ resultando no sinal janelado $a(t)$. Nos três gráficos inferiores: divisão da autocorrelação do sinal janelado $r_a(\tau)$ pela autocorrelação da janela $r_w(\tau)$ resultando na autocorrelação do quadro $r_x(\tau)$.



Fonte: Adaptado de (BOERSMA, 1993)

Um aspecto importante deste método é a amostragem do sinal. Para um sinal que não contém frequências maiores do que f_{max}, podemos realizar a amostragem em intervalos regulares $\Delta t \leq \frac{1}{2f_{max}}$ sem perder nenhuma informação do sinal original. Essa propriedade da amostragem de sinais digitais é enunciada no Teorema de Nyquist (RABINER; SCHAFER, 2011, Seção 2.5). Para que uma frequência f esteja presente no sinal amostrado, é necessário, portanto, que ele seja amostrado com frequência $f_s \geq 2f$.

A divisão do sinal completo em quadros é também uma forma de amostragem, já que para cada quadro é estimada uma autocorrelação normalizada $r_x(\tau)$ e um valor de F0 correspondente a $\frac{1}{\tau_0}$, sendo τ_0 o atraso correspondente ao maior pico de $r_x(\tau)$. É importante ressaltar que o tamanho do quadro influencia no atraso τ máximo em que um pico de $r_x(\tau)$ pode existir, o que por sua vez influencia nos valores de F0 possíveis de serem estimados. Por exemplo, para estimação de valores de F0 mais baixos, são necessários quadros de maior duração.

Especificamente para o método descrito nesta seção (BOERSMA, 1993), alguns parâmetros de entrada são essenciais na estimação de F0. Os mais relevantes são os

seguintes:

- Período de amostragem: determina a distância temporal entre dois quadros consecutivos;
- F0 mínimo: determina o valor mínimo de F0 que poderá ser estimado, influenciando diretamente no tamanho de cada quadro;
- F0 máximo: determina o valor máximo de F0 que poderá ser detectado, devendo ter valor maior que F0 mínimo e menor do que a frequência de Nyquist ($2f_{max}$);
- Limiar de silêncio (L_s) e Limiar de vozeamento (L_v): são valores entre 0 e 1 que determinam se um pico local de $r_x(\tau)$ é considerado para a estimação de F0 ou não. Se o módulo do pico local for menor ou igual a L_s vezes o módulo do pico global do sinal inteiro (ao longo de todos os quadros), esse pico local é considerado como silêncio e não influencia na estimação de F0. Por outro lado, se o módulo do pico local for menor do que L_v, esse pico é considerado como desvozeado e não influencia na estimação de F0;
- Custo de oitava: este parâmetro favorece valores mais altos de F0 quanto maior ele for. Uma das razões para sua existência é que para um sinal perfeitamente periódico, todos os picos serão de igual intensidade, mas aquele de menor atraso é que deve ser escolhido para o cálculo de F0.

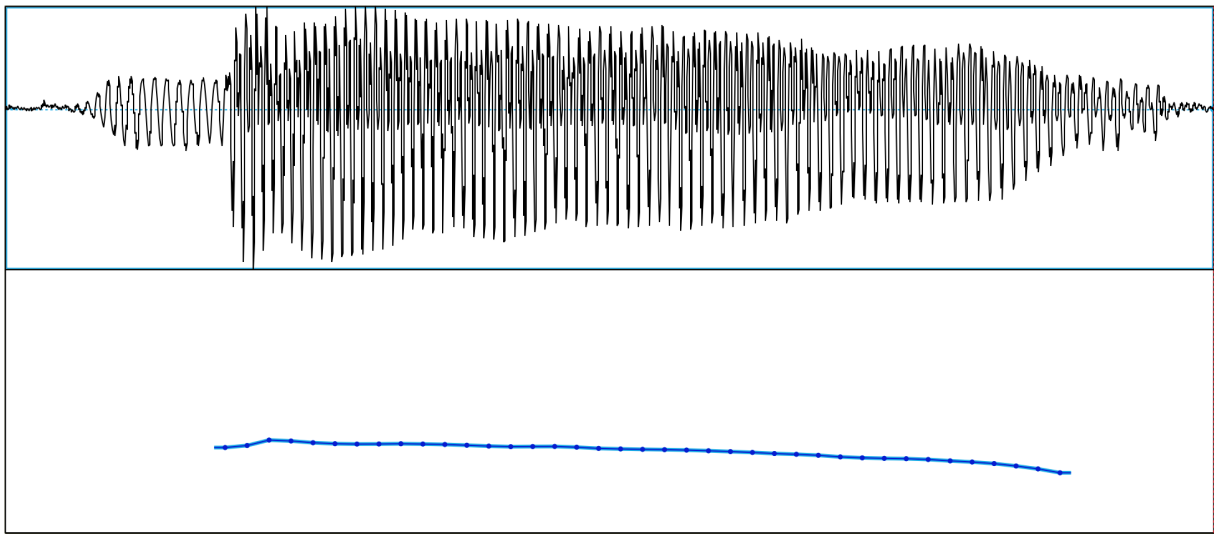
Neste trabalho, a implementação deste método que será utilizada será a do software de análise fonética Praat (BOERSMA; WEENINK, 2020). Além disso, o método é flexível e robusto, com resultados precisos para valores de F0 baixos (como 16Hz), médios (200Hz) e altos (1800Hz) (BOERSMA, 1993).

Cada palavra e cada sílaba gravadas tiveram suas curvas de F0 estimadas individualmente, com a adequação de alguns parâmetros para a minimização de artefatos, como por exemplo descontinuidades na curva de F0 e *pitch halving*, um fenômeno que acontece quando um valor de F0 é estimado com a metade de seu valor real. Os parâmetros alterados individualmente para cada palavra e cada sílaba foram 1) Limiar de silêncio, 2) Limiar de vozeamento, e 3) Custo de oitava.

Como o método utilizado para estimação de F0 é uma técnica de processamento de sinais de tempo curto, a forma de onda dos sinais acústicos foi dividida em quadros de curta duração processados individualmente. O tamanho de cada quadro é determinado pelo valor mínimo de F0 que queremos ser capazes de estimar, enquanto que a distância temporal entre quadros adjacentes é determinada pela frequência com a qual queremos obter um valor de F0. O valor utilizado como F0 mínimo foi igual a 75Hz, mais do que suficiente para falantes adultas do gênero feminino, que resulta em quadros com duração

de 40ms (BOERSMA, 1993). A distância temporal entre quadros adjacentes foi escolhida como 1/60s, de modo que um valor de F0 seja estimado a cada 1/60s, resultando numa frequência de amostragem de 60Hz para a curva de F0, igual às frequências de amostragem dos sinais de movimento capturadas pelo OPTOTRAK. Portanto, a partir dos sinais acústicos gravados durante o experimento a 44100Hz, foram extraídos sinais de F0 cuja dinâmica é diferente, sendo representado em 60Hz. A Figura 15 ilustra o a forma de onda de um sinal acústico e a curva de F0 estimada a partir dele.

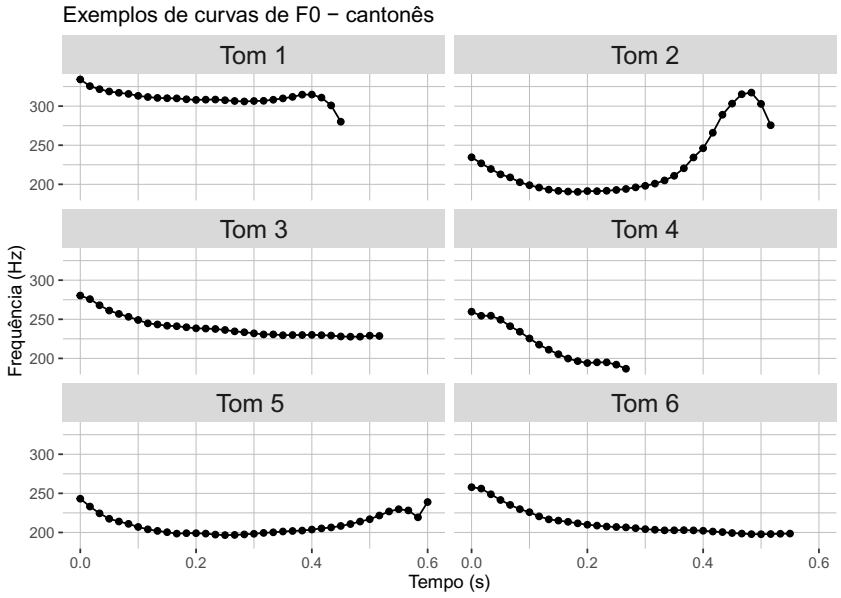
Figura 15 – Estimação de F0 por meio do software Praat (BOERSMA; WEENINK, 2020). No painel superior: a forma de onda do sinal acústico gravado pelo OPTOTRAK com frequência de amostragem igual a 44100Hz. No painel inferior: a curva de F0 estimada a partir da forma de onda do painel superior. Um valor de F0 é estimado a cada 1/60s, e cada valor é representado por um pequeno ponto.



Fonte: o autor

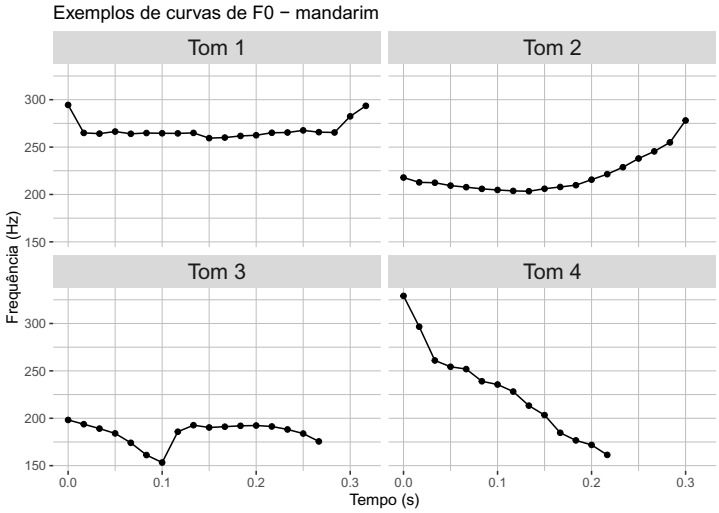
As Figuras 16, 17 e 18 trazem exemplos de curvas de F0 obtidas por meio do software Praat (BOERSMA; WEENINK, 2020) para cada um dos tons de cada uma das línguas estudadas: cantonês, mandarim e tailandês.

Figura 16 – Exemplos de curvas de F0 para cada um dos seis tons do cantonês. As curvas abaixo foram estimadas por meio do software Praat. O tom 1 é de nível alto (55), o tom 2 é crescente (25), o tom 3 é de nível médio (33), o tom 4 é descendente (21), o tom 5 é baixo-crescente (23) e o tom 6 é de nível baixo (22).



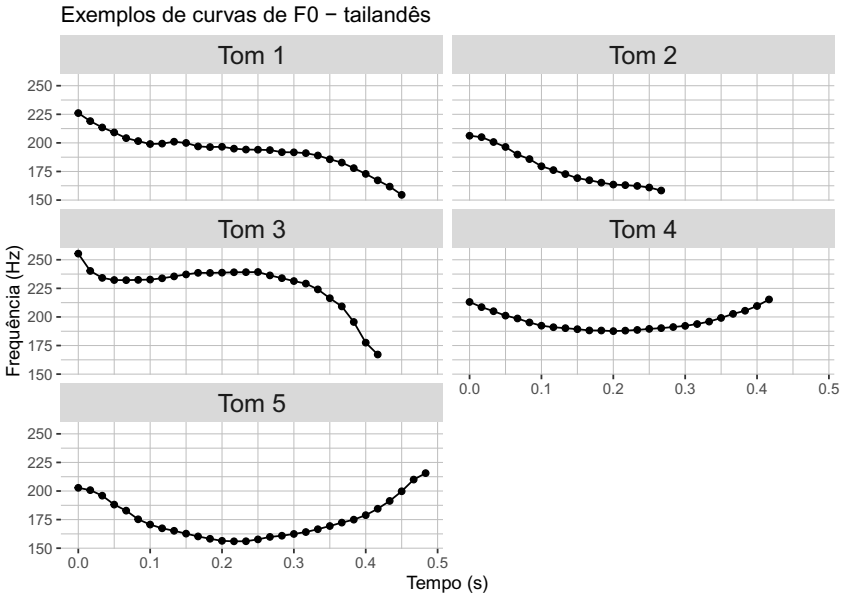
Fonte: o autor

Figura 17 – Exemplos de curvas de F0 para cada um dos quatro tons do mandarim. As curvas abaixo foram estimadas por meio do software Praat. O tom 1 é de nível alto (55), o tom 2 é crescente (35), o tom 3 é descendente-crescente (214) e o tom 4 é descendente (51).



Fonte: o autor

Figura 18 – Exemplos de curvas de F0 para cada um dos cinco tons do tailandês. As curvas abaixo foram estimadas por meio do software Praat. O tom 1 é descendente (HL), o tom 2 é baixo (L), o tom 3 é alto (H), o tom 4 é médio (M) e o tom 5 é crescente (LH).



Fonte: o autor

4.2.2 Processamento dos sinais visuais

As posições dos marcadores capturadas pelo OPTOTRAK durante o experimento consistem de duas componentes: uma associada com o movimento da face e outra associada com o movimento da cabeça. Por isso, o sinal das posições dos marcadores capturadas pelo OPTOTRAK será chamado ao longo deste trabalho de Movimento Total. Dado que este trabalho busca comparar as contribuições individuais dos movimentos da face e da cabeça na percepção de tons lexicais e também determinar movimentos específicos que são mais influentes nesse processo, precisamos separá-las. Isso é realizado por meio de um procedimento chamado de compensação do movimento da cabeça, que é essencialmente uma redução dos graus de liberdade do movimento da cabeça: a partir de 12 variáveis (coordenadas nos eixos (x, y, z) dos quatro marcadores de corpo rígido[2]) deseja-se descrever o movimento da cabeça por meio de apenas 6 variáveis (3 translações e 3 rotações nos eixos (x, y, z)).

Após a estimação do movimento da cabeça, é estimado o movimento da face. A estimação desses dois movimentos a partir das medições do OPTOTRAK é realizada da

[2] No caso deste trabalho, são os marcadores 1, 2, 3 e 4, cujo movimento depende exclusivamente da cabeça, já que estão presos num capacete utilizado pelas participantes (ver Figura 11).

seguinte forma: um novo sistema de coordenadas é estimado pelos marcadores 1, 2, 3 e 4, cujo movimento depende exclusivamente da cabeça, já que estão presos num capacete utilizado pelas participantes (ver Figura 11).

O movimento da cabeça em termos de 3 translações $\boldsymbol{T} = (t_1, t_2, t_3)$ e 3 rotações $\boldsymbol{R} = (r_1, r_2, r_3)$ é estimado a partir de um problema de mínimos quadrados não-linear que visa resolver o seguinte problema de minimização:

$$\min_{\boldsymbol{T},\boldsymbol{R}} F(\boldsymbol{T}, \boldsymbol{R}) = ||\boldsymbol{P} - \hat{\boldsymbol{P}}||_2, \tag{4.6}$$

sendo $||\boldsymbol{P} - \hat{\boldsymbol{P}}||_2$ a soma dos erros quadrados entre $\boldsymbol{P}$ e $\hat{\boldsymbol{P}}$. $\boldsymbol{P}$ é uma matriz de dimensão $3 \times N_{cr}$ com as posições nas coordenadas (x, y, z) dos N_{cr} marcadores de corpo rígido e $\hat{\boldsymbol{P}}$ a estimativa da matriz $\boldsymbol{P}$ realizada a partir dos parâmetros $\boldsymbol{T}$ e $\boldsymbol{R}$ a serem otimizados. Esse é um procedimento matemático de otimização iterativo cuja resolução mais detalhada pode ser vista em (COLEMAN; LI, 1996).

Este trabalho detalhará a obtenção da estimativa $\hat{\boldsymbol{P}}$, que é obtida da seguinte maneira:

1. Primeiramente é calculada a posição média de cada uma das três coordenadas de cada um dos N_{cr} marcadores de corpo rígido ao longo de todas as observações, resultando em $3 \times N_{cr}$ valores, organizados numa matriz $\boldsymbol{P_0}$ de dimensão $3 \times N_{cr}$, com as linhas correspondendo às coordenadas (x, y, z) e as colunas aos marcadores;

2. É aplicada à matriz de posições $\boldsymbol{P_0}$ a equação de movimento de corpo rígido com os parâmetros $\boldsymbol{T} = (t_1, t_2, t_3)$ e $\boldsymbol{R} = (r_1, r_2, r_3)$ a serem otimizados (os valores iniciais de $\boldsymbol{T}$ e $\boldsymbol{R}$ são vetores nulos de dimensão 1×3):

$$\hat{\boldsymbol{P}} = \boldsymbol{M} \times (\boldsymbol{P_0} - \boldsymbol{C_r}) + \boldsymbol{C_r} + \boldsymbol{T}, \tag{4.7}$$

sendo C_r o centro de rotação, que é o ponto médio de todos os N^3 marcadores ao longo de todas as observações, e $\boldsymbol{M}$ a matriz com as rotações do movimento dada por

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos r_1 & -\sin r_1 \\ 0 & \sin r_1 & \cos r_1 \end{bmatrix} \times \begin{bmatrix} \cos r_2 & 0 & -\sin r_2 \\ 0 & 1 & 0 \\ \sin r_2 & 0 & \cos r_2 \end{bmatrix} \times \begin{bmatrix} \cos r_3 & -\sin r_3 & 0 \\ \sin r_3 & \cos r_3 & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{4.8}$$

Deste modo, tem-se um problema de otimização que busca determinar, para cada amostra, o conjunto de parâmetros $\boldsymbol{T}$ e $\boldsymbol{R}$ que minimiza a Equação 4.6.

Após a estimativa do movimento da cabeça, descrito em cada amostra por três translações $\boldsymbol{T} = (t_1, t_2, t_3)$ e três rotações $\boldsymbol{R} = (r_1, r_2, r_3)$, é estimado o movimento da face

[3] No caso deste trabalho, $N = 33$ marcadores, de acordo com a Figura 11.

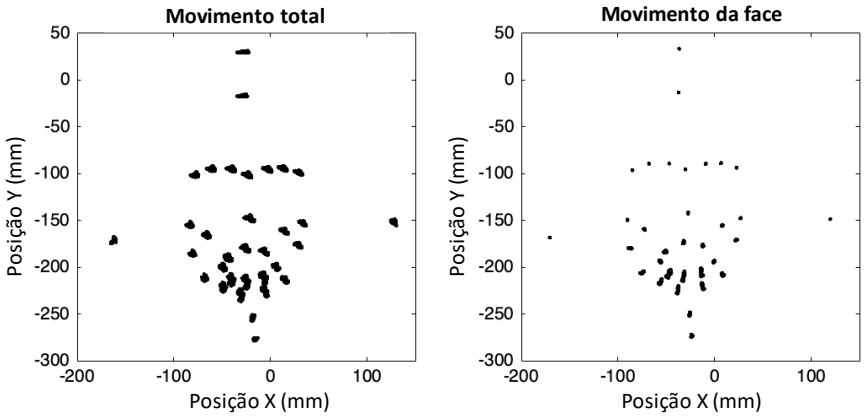
$\boldsymbol{P_{face}}$, em forma de matriz de dimensão $3 \times (N - N_{cr})$, com as linhas correspondendo às coordenadas (x, y, z) e as colunas correspondendo aos $(N - N_{cr})$ marcadores que não são de corpo rígido.

Para a estimativa de $\boldsymbol{P_{face}}$ são utilizadas, para cada amostra, os valores de $\boldsymbol{T}$ e $\boldsymbol{R}$ correspondentes estimados anteriormente. A estimativa é realizada aplicando a equação de movimento de corpo rígido, mas de uma forma diferente daquela descrita na Equação 4.7, com a posição inicial $\boldsymbol{P_{0m}}$ sendo a posição absoluta dos marcadores em cada amostra:

$$\boldsymbol{P_{face}} = \boldsymbol{M'} \times (\boldsymbol{P_{0m}} - \boldsymbol{C_r} - \boldsymbol{T}) + \boldsymbol{C_r}, \tag{4.9}$$

Deste modo, a Equação 4.9 ignora a rotação e a translação do movimento da cabeça que estão presentes no movimento dos marcadores $\boldsymbol{P_{0m}}$ e resulta no movimento da face $\boldsymbol{P_{face}}$. A Figura 19 ilustra a diferença entre os movimentos dos marcadores antes e depois do procedimento de compensação do movimento da cabeça.

Figura 19 – Movimento dos marcadores ao longo de 2000 amostras (33, 33s). À esquerda: movimento original capturado pelo OPTOTRAK composto pelas componentes da face e da cabeça. À direita: componente de movimento da face após a realização da compensação do movimento da cabeça. A ausência do movimento da cabeça é perceptível no painel da direita.



Fonte: o autor

Com a estimativa dos movimentos da cabeça e da face, temos mais dois sinais de movimentos disponíveis. A Tabela 11 sumariza esses sinais de movimento e as suas dimensões. Nota-se que a partir dos 99 sinais do Movimento Total foram gerados 87+6 = 93 sinais, pois houve uma redução dos graus de liberdade do movimento da cabeça de 12 para 6.

Tabela 11 – Sumarização dos sinais de movimento.

Tipo de sinal	**Composição**
Movimento Total	99 sinais: 33 marcadores × 3 coordenadas (x, y, z)
Face	87 sinais: 29 marcadores × 3 coordenadas (x, y, z)
Cabeça	6 sinais: 3 translações (x, y, z) + 3 rotações (x, y, z)

4.3 Aproximação polinomial

Após os procedimentos de pré-processamento específicos para cada tipo de sinal, descritos nas Seções 4.2.1 e 4.2.2, estão disponíveis os seguintes sinais, descritos na Tabela 12.

Os dados de todos os tipos de sinal foram divididos em palavras e/ou sílabas (chamadas de elocuções deste ponto em diante) individuais durante a gravação, cada uma delas com uma duração diferente. Essa diferença de durações é um problema, pois para realizar classificação estatística com base em todas as gravações é necessário que todos os vetores de entrada tenham o mesmo comprimento. Resolvemos esse problema parametrizando cada um desses sinais por meio de aproximações polinomiais. Dois sinais de comprimento diferentes podem ser descritos pelo mesmo número $p + 1$ de coeficientes, caso sejam ambos aproximados por polinômios da mesma ordem p.

As aproximações polinomiais foram realizadas individualmente para cada tipo de sinal e para cada elocução por meio da decomposição QR (HORN; JOHNSON, 1985, Seção 2.6). Por exemplo, cada um dos 99 sinais do Movimento Total foram aproximados individualmente para cada uma das elocuções gravadas. Antes da aproximação, cada sinal foi centralizado na coordenada $x = 0$. Isso foi feito de modo a obter maior uniformidade entre os polinômios obtidos, já que o mesmo sinal pode resultar em aproximações polinomiais diferentes a depender do seu ponto de origem no tempo.

Embora o procedimento de aproximação polinomial seja capaz de descrever sinais de comprimentos diferentes pelo mesmo número de coeficientes, ele introduz erro, pois as aproximações nem sempre são capazes de descrever perfeitamente o sinal original. Uma forma de medir esse erro é por meio do erro médio quadrático (MSE, do inglês *Mean*

Tabela 12 – Sumarização dos sinais disponíveis após a aplicação de procedimentos de pré-processamento específicos para os sinais de dois tipos: de movimento e acústicos.

Tipo de sinal	**Composição**
Movimento Total	99 sinais: 33 marcadores × 3 coordenadas (x, y, z)
Face	87 sinais: 29 marcadores × 3 coordenadas (x, y, z)
Cabeça	6 sinais: 3 translações (x, y, z) + 3 rotações (x, y, z)
Curva de F0	1 sinal: valores de F0 estimados a cada $1/60s$

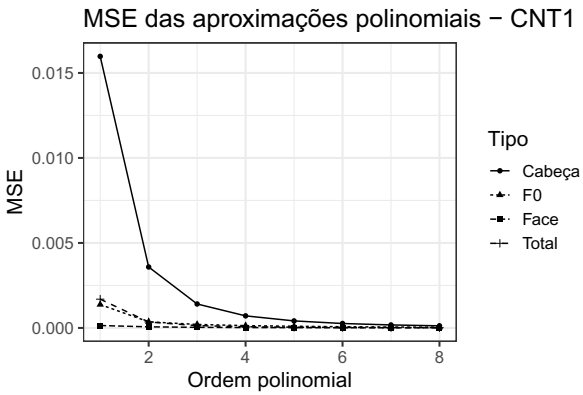
squared error), definido por

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2 \tag{4.10}$$

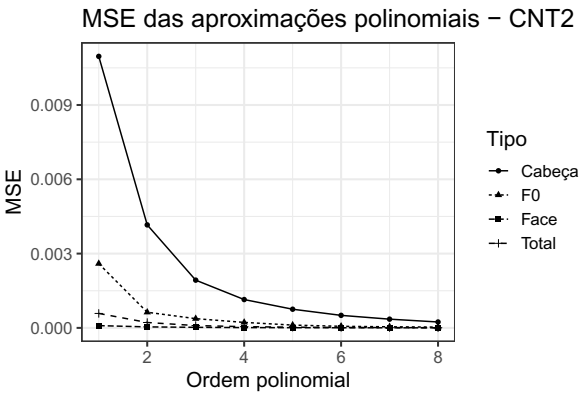
sendo n o número de observações, y_i o valor real e $\hat{y}_i = \hat{f}(x_i)$ seu valor predito. Um MSE será pequeno se as respostas preditas forem consistentemente próximas das respostas reais (JAMES et al., 2013, Seção 2.2).

Um fator determinante para a aproximação polinomial é a ordem p: quanto maior a ordem, maior o grau de liberdade do polinômio e, consequentemente, maior sua capacidade de aproximar sinais. As Figuras 20, 21 e 22 trazem os MSEs das aproximações polinomiais de ordem $p = 1$ até $p = 8$ obtidos para cada um dos tipos de sinal para cada uma das participantes. Para cada tipo de sinal, um valor de MSE é obtido para cada ordem polinomial da seguinte maneira: 1) para cada elocução é obtido o MSE médio entre todos os sinais (99 para Movimento Total, 87 para Face, 6 para Cabeça e 1 para F0); 2) para cada ordem polinomial é obtido o MSE médio entre todas as elocuções.

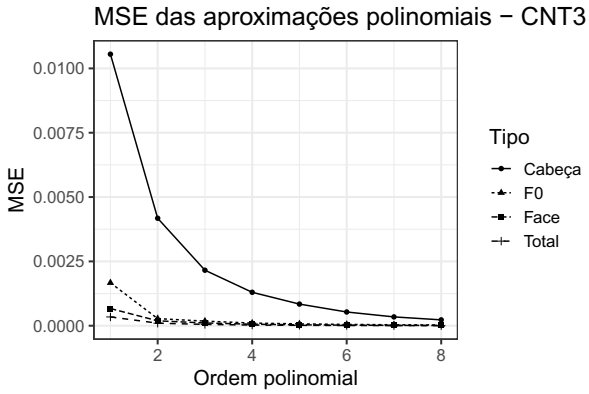
Figura 20 – MSE das aproximações polinomiais para cada tipo de sinal para cada participante falante nativa de cantonês. Cada valor é a média entre todos os sinais de cada elocução.



(a) Participante CNT1.



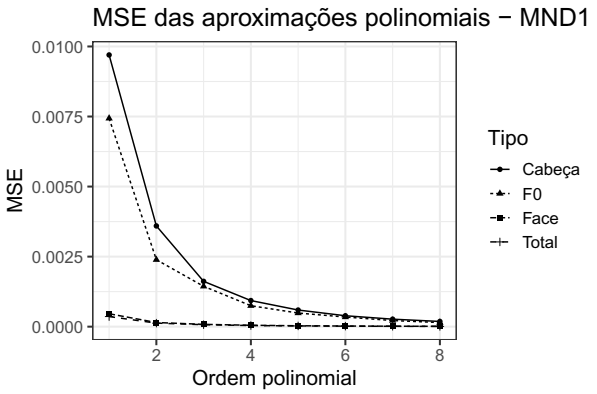
(b) Participante CNT2.



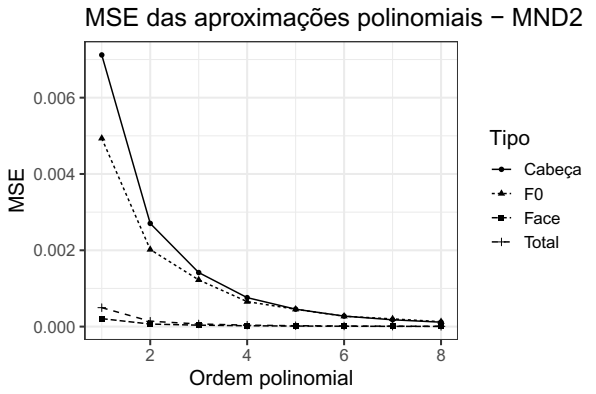
(c) Participante CNT3.

Fonte: o autor

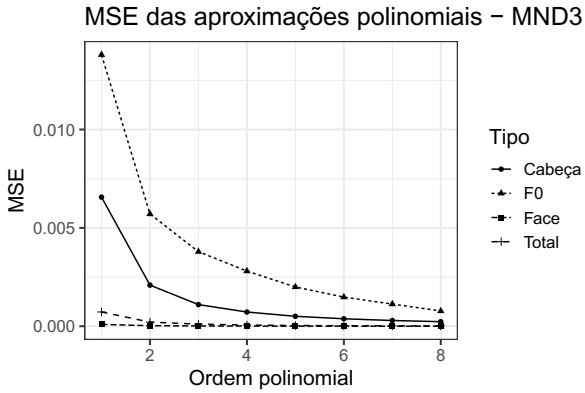
Figura 21 – MSE das aproximações polinomiais para cada tipo de sinal para cada participante falante nativa de mandarim. Cada valor é a média entre todos os sinais de cada elocução.



(a) Participante MND1.



(b) Participante MND2.

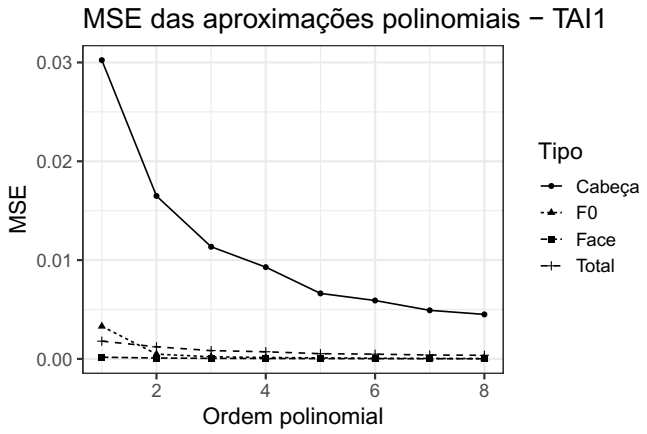


(c) Participante MND3.

Fonte: o autor

Com o aumento da ordem polinomial, mais coeficientes são necessários para aproximar o sinal, o que faz com que o erro diminua. Queremos utilizar uma mesma ordem polinomial p para todas as participantes de todas as línguas de forma a obter um equilíbrio entre baixo número de coeficientes e baixo erro. Com base nas Figuras 20, 21 e 22, temos que após a ordem $p = 3$ os valores de erro começam a decrescer mais lentamente do que decrescem antes de $p = 3$. Isso é perceptível principalmente para as participantes falantes nativas de cantonês e mandarim. Para a participante falante nativa de tailandês pode ser observada outra ordem na qual isso acontece: $p = 5$. Contudo, para manter a mesma ordem p para todas as participantes e para ter o menor número de coeficientes possível, escolhemos a ordem $p = 3$ para ser utilizada ao longo das análises deste trabalho. Um polinômio de ordem $p = 3$ é descrito por 4 coeficientes e possui dois pontos de inflexão ao longo de sua trajetória (um polinômio de ordem $p = 1$ é uma reta, que mantém uma única direção, e um polinômio de ordem $p = 2$ é uma parábola, que realiza uma mudança de

Figura 22 – MSE das aproximações polinomiais para cada tipo de sinal para a participante falante nativa de tailandês RP. Cada valor é a média entre todos os sinais de cada elocução.



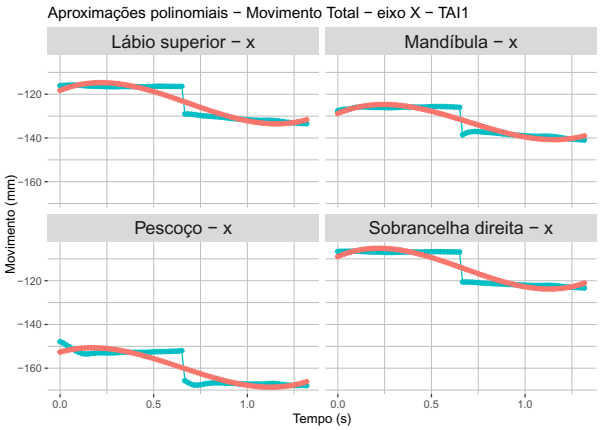
Fonte: o autor

direção, por exemplo). As Figuras 23, 24, 25 e 26 ilustram as aproximações obtidas com polinômios de ordem $p = 3$ para os diferentes tipos de sinal.

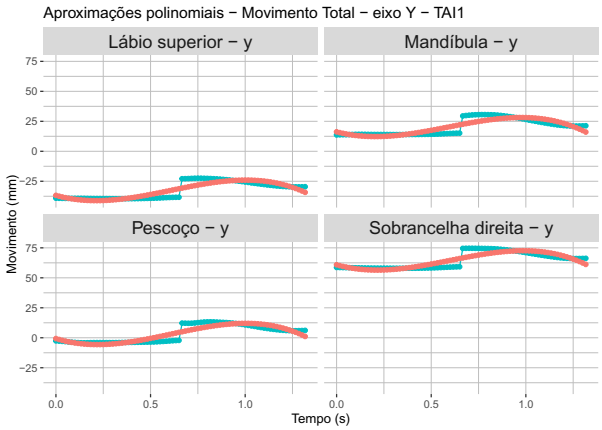
As Figuras 23 a 26 ilustram as características de cada sinal utilizado neste trabalho. Os sinais de Movimento Total e Cabeça sofrem mudanças rápidas de posição, provavelmente devidas a problemas ocorridos no processo de captura da posição dos marcadores, representadas pelas mudanças bruscas de nível entre amostras adjacentes nas Figuras 23 e 25. Essas mudanças rápidas de posição são artefatos e aumentam o erro da aproximação polinomial para esses sinais, pois os polinômios de ordem $p = 3$ não são capazes de aproximar essas mudanças ao mesmo tempo que aproximam o restante do sinal. Esses artefatos são também observados nos sinais de movimento da face, mas com menos amplitude. Por outro lado, os sinais de F0 não apresentam esses artefatos, mas sim pequenas variações que a aproximação polinomial de ordem $p = 3$ é capaz de aproximar sem introduzir um erro muito alto.

É importante ressaltar que as Figuras 23 a 26 trazem as aproximações apenas para uma participante, e não retratam um caso geral. Contudo, a tendência de que os sinais de movimento da cabeça são aproximados com maior erro é observada para todas as participantes (à exceção da participante MND3, na Figura 21c, em que o sinal de $F0$ é aproximado com maior erro), o que provavelmente se deve às mudanças bruscas de posição ilustradas pela Figura 25.

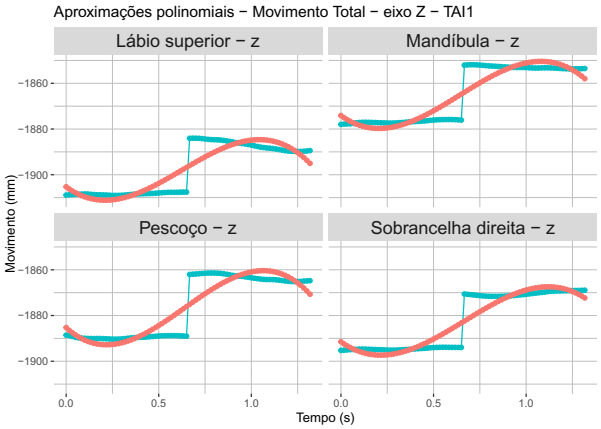
Figura 23 – Aproximações polinomiais de ordem $p = 3$ de alguns sinais de Movimento Total. Os sinais foram produzidos pela participante TAI1, falante nativa de tailandês. Em azul estão os sinais originais e em vermelho as aproximações polinomiais.



(a) Movimento no eixo X.



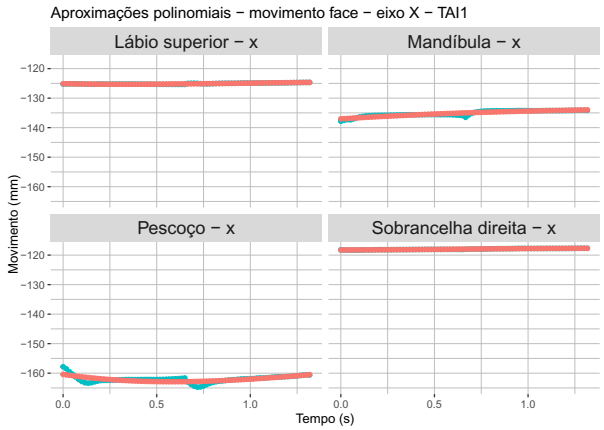
(b) Movimento no eixo Y.



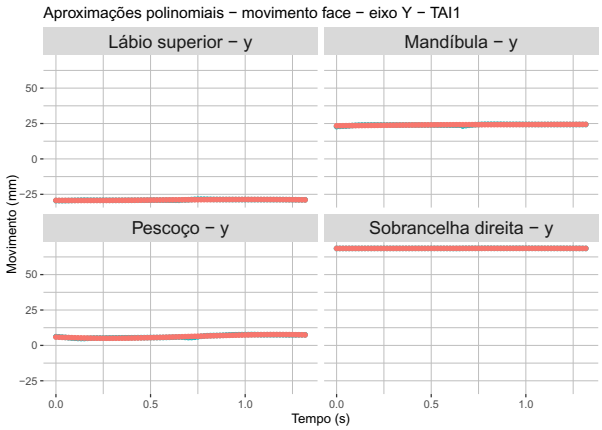
(c) Movimento no eixo Z.

Fonte: o autor

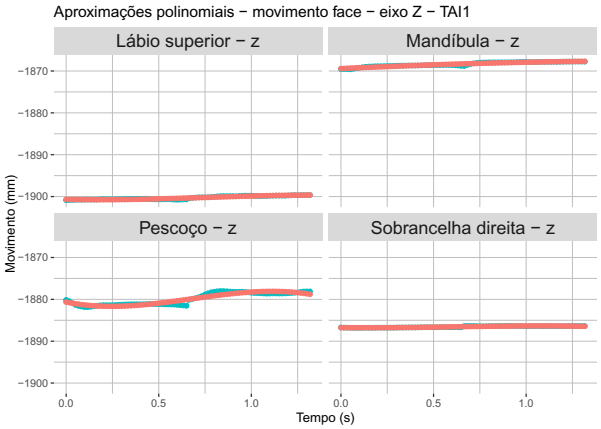
Figura 24 – Aproximações polinomiais de ordem $p = 3$ de alguns sinais de movimento da face. Os sinais foram produzidos pela participante TAI1, falante nativa de tailandês. Em azul estão os sinais originais e em vermelho as aproximações polinomiais.



(a) Movimento no eixo X.



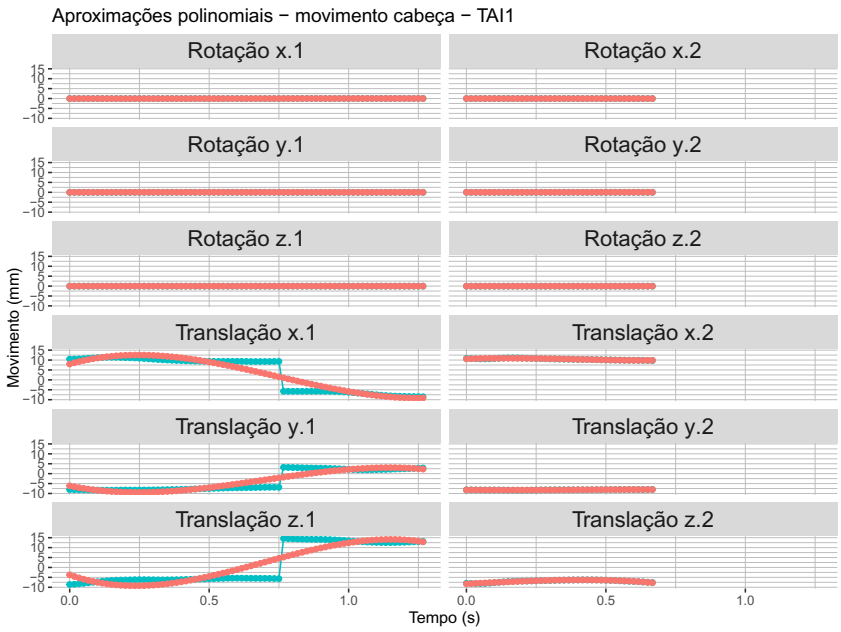
(b) Movimento no eixo Y.



(c) Movimento no eixo Z.

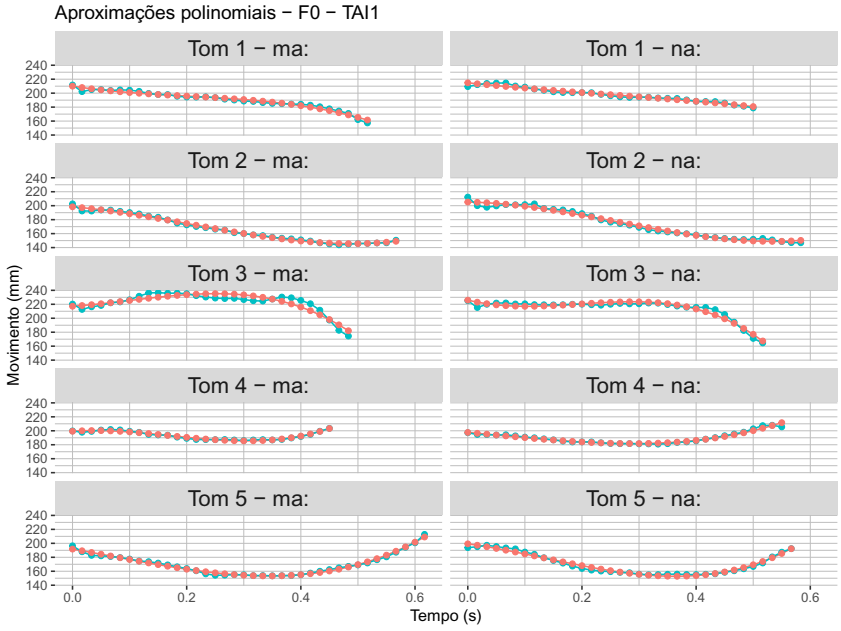
Fonte: o autor

Figura 25 – Aproximações polinomiais de ordem $p = 3$ dos sinais de movimento da cabeça. Os sinais foram produzidos pela participante TAI1, falante nativa de tailandês. Em azul estão os sinais originais e em vermelho as aproximações polinomiais.



Fonte: o autor

Figura 26 – Aproximações polinomiais de ordem $p = 3$ dos sinais de F0. Os sinais foram produzidos pela participante TAI1, falante nativa de tailandês. Em azul estão os sinais originais e em vermelho as aproximações polinomiais.



Fonte: o autor

5 Método de classificação

Para responder as perguntas do trabalho expostas no Capítulo 1, serão utilizados métodos de classificação estatística que, com base nos diferentes tipos de sinais da base de dados exposta no Capítulo 4, realizarão a tarefa de classificar entre diferentes tons lexicais de uma mesma língua. Este capítulo apresentará brevemente conceitos básicos sobre classificação estatística e, em seguida, o formato no qual a base de dados será utilizada como entrada nos classificadores. Em seguida serão apresentados os quatro métodos de classificação utilizados e, por fim, a forma de validação dos resultados.

5.1 Classificação estatística

Seja o seguinte exemplo: uma universidade deseja avaliar a produção acadêmica do seu corpo docente e, para isso, levantou as seguintes informações de cada docente: tempo de carreira, produção acadêmica total e titulação. O intuito dessa avaliação é desenvolver metas de produção acadêmica a serem seguidas ao longo da carreira pelos grupos de docentes com diferentes titulações: mestrado, doutorado e pós-doutorado.

Nesse contexto, os tempos de carreira de cada docente são as **variáveis de entrada** e a produção acadêmica de cada docente, dividida por titulação, a **variável de saída**. Variáveis de entrada são geralmente denotadas por X, com algum índice subscrito que possa distingui-las das demais. No exemplo acima, X_1 pode ser os tempos de carreira dos docentes com mestrado, X_2 os tempos de carreira dos docentes com doutorado e X_3 os tempos de carreira dos docentes com pós-doutorado. As variáveis de entrada são também comumente chamadas de preditores, variáveis independentes, características ou simplesmente variáveis. A variável de saída é comumente chamada de variável dependente ou de resposta e denotada por Y (JAMES et al., 2013, Seção 2.1).

Supondo que os preditores X_i influenciem a resposta Y, pode-se escrever essa relação como

$$Y = f(X) + \epsilon \tag{5.1}$$

sendo ϵ um erro aleatório, independente de X e com média zero e f uma função fixa, mas ainda desconhecida, de X. Na Equação 5.1, f representa a informação sistemática a respeito de Y que pode ser obtida a partir de X. Aprendizado estatístico pode ser definido, então, como um conjunto de abordagens para estimação de f (JAMES et al., 2013, Seção 2.1). Dentro do que se chama de aprendizado estatístico, existem problemas que trabalham com a resposta Y quantitativa, chamados de regressão, e problemas que trabalham com

a resposta Y qualitativa, chamados de classificação. De interesse para este trabalho é o problema de classificação entre classes qualitativas, os tons lexicais de cantonês, mandarim e tailandês.

O resultado obtido em problemas de classificação é geralmente definido como a acurácia, que pode ser definida como a porcentagem de classificações corretas. O complemento da acurácia, o erro, também pode ser utilizado. Disto surge outra questão: quais dados que o classificador vai classificar? Todo classificador estatístico necessita de um conjunto de preditores X associado a uma resposta Y para que possa encontrar os padrões nos dados relevantes para a classificação. A esses dados dá-se o nome de conjunto de treinamento. Quando a acurácia do classificador é medida com os mesmos dados utilizados para treiná-lo, é obtida a acurácia de treinamento. Em situações reais, contudo, a acurácia de treinamento não é tão importante quanto a acurácia de teste, pois se deseja saber a precisão do classificador com dados de entrada previamente desconhecidos, ou seja, dados que não foram utilizados em seu treinamento. Um exemplo disso é um classificador cujos preditores X são parâmetros sanguíneos de pacientes de um dado hospital ou região e cuja resposta Y qualitativa é se o paciente possui ou não diabetes. Na prática, esse classificador é treinado com os dados de pacientes passados ou presentes que já possuem diagnóstico, mas deve prever com precisão o diagnóstico de diabetes para pacientes futuros, cujos parâmetros não foram utilizadas em seu treinamento. Deste modo, a acurácia de teste é uma medida mais rigorosa da acurácia do classificador do que a acurácia de treinamento.

Modelos de classificação estatística podem apresentar grandes diferenças entre suas acurácias de treinamento e de teste. Uma situação comumente encontrada em modelos que apresentam acurácia de treinamento alta e acurácia de teste baixa é o sobreajuste (do inglês *overfitting*). O sobreajuste ocorre quando o modelo encontra padrões demais nos dados de treinamento, podendo alguns desses padrões serem causados pelo acaso e não pelas propriedades da distribuição real dos dados (JAMES et al., 2013, Seção 2.2.1), ou seja, o modelo é muito adequado aos dados de treinamento, mas com isso perde flexibilidade, falhando em se adequar a dados de teste. Uma solução para casos em que o sobreajuste ocorre é a utilização de modelos mais simples, que tem menor capacidade de adaptação aos dados de treinamento, gerando resultados menos precisos (com maior viés), mas também menos variáveis (com menor variância).

Cabe aqui explicar os termos viés e variância. De forma simples, viés se refere ao erro que ocorre quando se tenta descrever um problema complexo com um modelo simples enquanto que variância se refere ao tanto que um modelo varia caso seja treinado por um conjunto diferente de dados. Um modelo com viés mínimo e variância máxima seria um curva que passa por todos os pontos y_i da saída Y de treinamento. Por outro lado, um modelo com viés máximo e variância mínima seria uma linha horizontal (JAMES et al., 2013, Seção 2.2.2).

5.2 Formato da entrada

A base de dados deste trabalho é composta por sinais audiovisuais, que são os preditores X dos classificadores. Os tons lexicais associados a cada elocução compõem a resposta Y. Como descrito no Capítulo 4, os dados acústicos e visuais adquiridos foram parametrizados por meio de coeficientes que descrevem aproximações polinomiais de ordem $p = 3$. Esse sinais devem estar num formato apropriado para servirem de entrada para os métodos de classificação, que geralmente lidam com dados na seguinte estrutura: matrizes em que cada linha corresponde a uma observação e cada coluna corresponde a uma variável ou dimensão dessa observação. Na base de dados utilizada, cada elocução é uma observação.

Os diferentes tipos de sinais (F0, Movimento Total, Face e Cabeça) possuem dimensões diferentes e foram arranjados em matrizes diferentes, já que os classificadores processarão cada tipo de sinal individualmente. Sendo assim, a organização dos dados em matrizes apropriadas para os classificadores é realizada da seguinte forma: seja N o número de dimensões do sinal ($N = 99$ para o sinal Movimento Total, $N = 87$ para o sinal Face, $N = 6$ para o sinal Cabeça e $N = 1$ para o sinal F0). Após a aproximação polinomial, todas as elocuções são descritas por 4 coeficientes por dimensão, independentemente da duração original. Logo, as M elocuções diferentes de cada participante (CNT1, CNT2, CNT3, MND1, MND2, MND3 e TAI1) são matrizes de dimensão $4 \times N$.

Para que cada elocução seja descrita por uma única linha, é realizado então o procedimento de vetorização, que modifica a dimensão de cada elocução para $1 \times (4N)$. Sendo a matriz original A

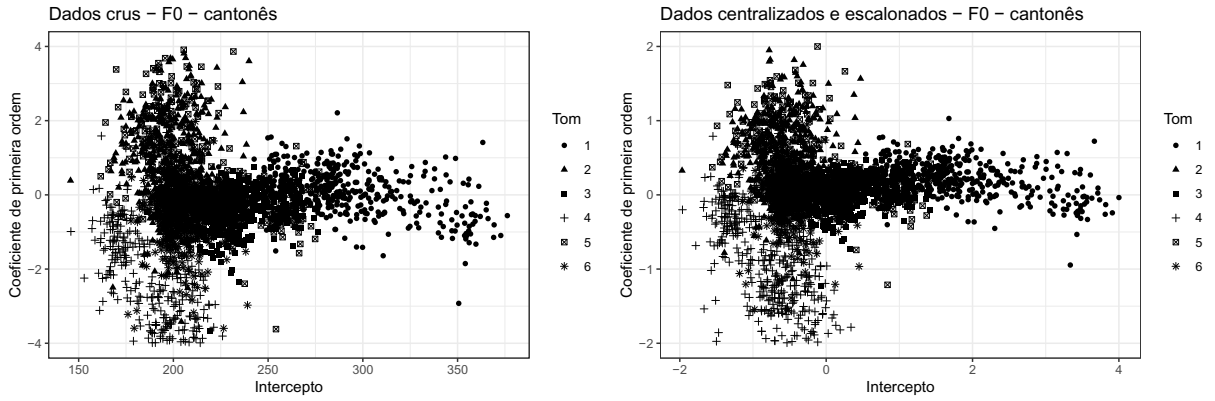
$$A = \begin{bmatrix} a_{11} & \dots & a_{1N} \\ a_{21} & \dots & a_{2N} \\ a_{31} & \dots & a_{3N} \\ a_{41} & \dots & a_{4N} \end{bmatrix}, \tag{5.2}$$

sua versão vetorizada é

$$A_V = [a_{11} \ \dots \ a_{1N} \, a_{21} \ \dots \ a_{2N} \, a_{31} \ \dots \ a_{3N} \, a_{41} \ \dots \ a_{4N}]. \tag{5.3}$$

A matriz X, que servirá como entrada para os classificadores é então de dimensão $M \times (4N)$, com as M linhas correspondendo às elocuções e as $4N$ colunas às variáveis ou dimensões de cada observação. Por fim, antes de servir de entrada para o classificador, cada matriz X é centralizada e escalonada, de modo que tenha média nula e desvio-padrão unitário. A centralização é realizada de modo a normalizar a faixa de valores que cada sinal ocupa, para que algumas variáveis com valores médios muito maiores que outras não dominem a classificação e o escalonamento (do inglês *scaling*) é feito de forma a normalizar as flutuações de valor que variáveis diferentes sofrem. As Figuras 27b, 28b e 29b

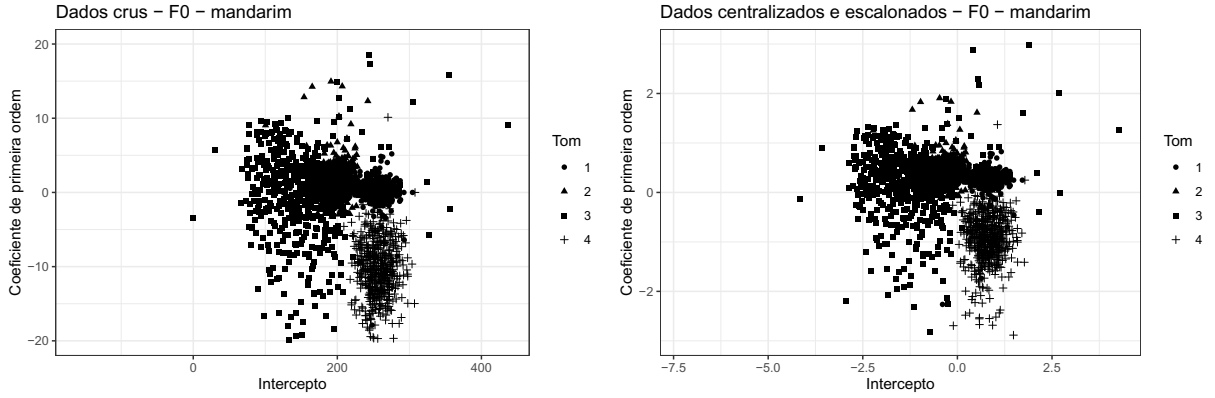
Figura 27 – Dois coeficientes do sinal de F0 para cada uma das elocuções de cantonês. À esquerda, os coeficientes antes das operações de centralização e escalonamento. À direita, os coeficientes após as operações de centralização e escalonamento.



(a) Dados crus. (b) Dados centralizados e escalonados.

Fonte: o autor

Figura 28 – Dois coeficientes do sinal de F0 para cada uma das elocuções de mandarim. À esquerda, os coeficientes antes das operações de centralização e escalonamento. À direita, os coeficientes após as operações de centralização e escalonamento.



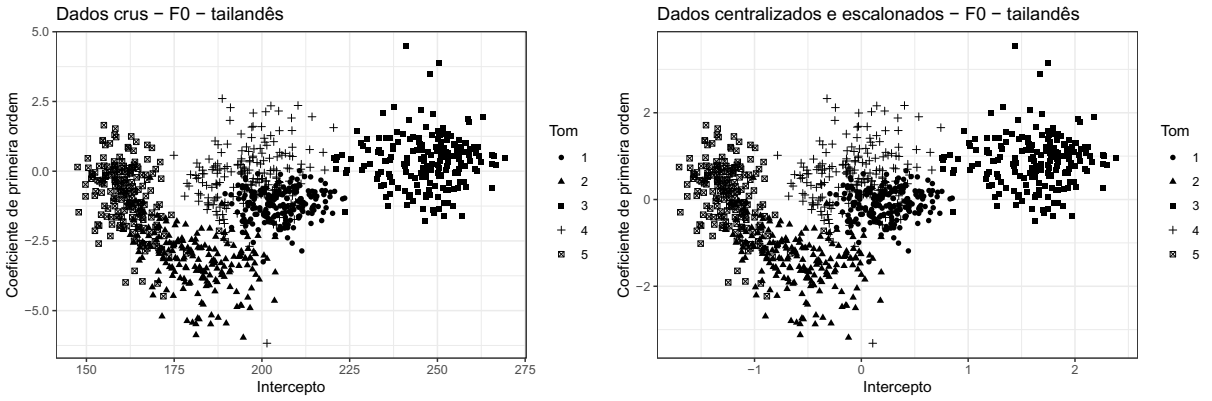
(a) Dados crus. (b) Dados centralizados e escalonados.

Fonte: o autor

mostram, para as três línguas, os dois primeiros coeficientes do sinal de $F0$ (o intercepto e o coeficiente de primeira ordem) antes e depois de serem centralizados e escalonados. Há diferenças nos valores dos eixos antes e após os procedimentos, principalmente no intercepto, que possui valores originais de maior módulo, mas a distribuição das produções se manteve semelhante.

A matriz Y é composta pelos tons lexicais associados a cada uma das M elocuções de cada participante e possui dimensão $M \times 1$.

Figura 29 – Dois coeficientes do sinal de F0 para cada uma das elocuções de tailandês. À esquerda, os coeficientes antes das operações de centralização e escalonamento. À direita, os coeficientes após as operações de centralização e escalonamento.



(a) Dados crus.

(b) Dados centralizados e escalonados.

Fonte: o autor

5.3 Métodos utilizados

Neste trabalho serão utilizados quatro métodos diferentes de classificação estatística. Dois deles são métodos lineares: a Análise Discriminante Linear (LDA) e a Máquina de Vetores de Suporte (SVM) com *kernel* linear; e dois deles são métodos não-lineares: o K-vizinho mais próximos (KNN) e a Máquina de Vetores de Suporte (SVM) com *kernel* radial. A principal distinção entre os métodos de classificação lineares e não-lineares é sua capacidade de traçar limites entre as classes. Enquanto que nos métodos lineares eles são definidos por funções lineares, nos métodos não-lineares eles são definidos por funções mais complexas. A seguir são descritos os quatro métodos de classificação utilizados.

5.3.1 Análise Discriminante Linear (LDA)

A Análise Discriminante Linear (LDA, do inglês *Linear Discriminant Analysis*) é uma técnica de classificação que consiste basicamente de duas etapas: 1) a projeção dos preditores X num espaço de dimensão reduzida e 2) a determinação de funções lineares que separam as diferentes classes presentes na resposta Y.

Sua capacidade de projetar X num espaço de dimensão reduzida faz com que a LDA possa ser utilizada de forma parcial, apenas para reduzir a dimensionalidade dos dados, de forma similar a técnicas como a Análise de Componentes Principais (PCA) (JAMES et al., 2013, Seção 10.2). Se por um lado a redução de dimensionalidade da PCA é realizada sem o conhecimento de Y e de forma a maximizar a variância de X em cada uma das componentes principais, a LDA reduz a dimensionalidade de X de forma supervisionada (com o conhecimento de Y, as classes associadas a cada observação de X)

e de forma a maximizar a separabilidade entre as classes (BISHOP, 2006, Seção 4.1).

Isso é realizado por meio da determinação de uma matriz de rotação $\boldsymbol{W}$ que projeta os dados X em novas dimensões, não necessariamente ortogonais entre si. A matriz de rotação $\boldsymbol{W}$ é obtida a partir das matrizes de covariância intra-classes ($\boldsymbol{S_W}$) e entre-classes ($\boldsymbol{S_B}$), dadas pelas equações

$$\boldsymbol{S_W} = \sum_{c=1}^{N_c} \sum_{n \in C_c} (\boldsymbol{x}_n - \boldsymbol{m}_c)(\boldsymbol{x}_n - \boldsymbol{m}_c)^T e \tag{5.4}$$

$$\boldsymbol{S_B} = \sum_{c=1}^{N_c} M_c (\boldsymbol{m}_c - \boldsymbol{m})(\boldsymbol{m}_c - \boldsymbol{m})^T, \tag{5.5}$$

sendo N_c o número de classes, C_c o conjunto de observações correspondentes à classe c, $\boldsymbol{x}_n$ uma observação multidimensional, $\boldsymbol{m}_c$ a observação média da classe c, $\boldsymbol{m}$ a média de todas as observações e M_c o número de observações pertencentes à classe c. A matriz de rotação é então obtida por

$$\boldsymbol{W} = \boldsymbol{S_W}^{-1} \boldsymbol{S_B}, \tag{5.6}$$

sendo o número máximo de dimensões dessa matriz igual a $N_c - 1$, sendo N_c o número de classes diferentes presentes em Y (BISHOP, 2006, Seção 4.1).

Com os preditores X agora projetados em dimensões que maximizam a separação entre classes diferentes, a LDA define hiperplanos que definem os limites entre as N_c classes. Para isso, são feitos dois importantes pressupostos (JAMES et al., 2013, Seção 4.4):

- As distribuições $f_c(X)$ de cada classe c são Gaussianas multidimensionais;
- Todas as classes possuem a mesma matriz de covariância Σ.

Desta forma, para cada par de classes j e l é definida uma função linear, chamada de discriminante linear, que serve como uma limiar de decisão, definindo os limites de classificação para cada classe:

$$x^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j - \frac{1}{2} \boldsymbol{\mu}_j^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_j + \log \pi_j = x^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l - \frac{1}{2} \boldsymbol{\mu}_l^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_l + \log \pi_l, \tag{5.7}$$

sendo $\boldsymbol{\mu}_j$ e $\boldsymbol{\mu}_l$ as médias das classes j e l, respectivamente, π_j e π_l as probabilidade *a priori*[1] das classes j e l, respectivamente e x a variável dependente que determina o limiar de classificação. A observação x_i é então classificada de acordo com sua posição relativa aos limiares de classificação, que definem regiões de pertencimento para cada classe.

[1] A probabilidade *a priori* π_j é probabilidade inicial de que uma observação pertença à classe j, dada por $\pi_j = n_j/n$, sendo n_j o número de observações que pertencem à classe j e n o número total de observações (JAMES et al., 2013, Seção 4.4.1).

5.3.2 K-vizinhos mais próximos (KNN)

O K-vizinhos mais próximos (KNN, do inglês *K-nearest neighbors*) é uma abordagem simples de classificação que é capaz, em muitos casos, de produzir resultados semelhantes ao classificador de Bayes, que é o classificador ideal, mas raramente viável de ser aplicado. Dado um número inteiro positivo K_{NN} e uma observação de teste x_0, o classificador KNN primeiramente identifica os K_{NN} pontos no conjunto de treinamento mais próximos de x_0, aqui denominados de N_0. Com base em N_0, um conjunto composto por K_{NN} pontos, o KNN estima a probabilidade condicional de que a observação x_0 pertença à classe j como sendo a fração dos pontos de N_0 que pertencem à j:

$$Pr\left(Y = j | X = x_0\right) = \frac{1}{K_{NN}} \sum_{i \in N_0} I(y_i = j), \tag{5.8}$$

sendo $I(y_i = j)$ uma variável indicadora que assume o valor $I(y_i = j) = 1$ caso $y_i = j$ e o valor $I(y_i = j) = 0$ caso $y_i \neq j$.

Por fim, o KNN aplica a regra de Bayes e classifica x_0 na classe j com a maior probabilidade (JAMES et al., 2013, Seção 2.2.3). O parâmetro K_{NN} é fundamental no desempenho do KNN. Para valores pequenos de K_{NN}, como por exemplo $K_{NN} = 1$, o classificador se torna muito flexível, com baixo viés e alta variância. Por outro lado, para valores altos de K_{NN}, como por exemplo $K_{NN} = 100$, o classificador se torna pouco flexível, aproximando-se de uma classificador linear, com alto viés e baixa variância.

5.3.3 Máquina de Vetores de Suporte

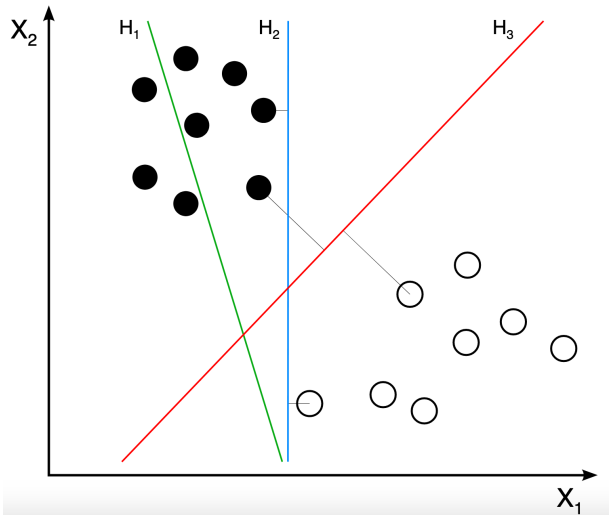
As máquinas de vetores de suporte (SVM, do inglês *Support vector machines*) são um conjunto de técnicas de classificação estatística que se baseiam em definir hiperplanos[2] de separação entre classes.

O hiperplano definido pela SVM é chamado de hiperplano de margem máxima, ou seja, o hiperplano que está o mais afastado das observações de treinamento possível. Um exemplo de diferentes hiperplanos é dado na Figura 30. A classificação é realizada por meio da posição da amostra de teste relativa ao hiperplano e quanto maior a distância dessa amostra de teste ao hiperplano, maior a certeza da classificação. As amostras de treinamento ligadas por linhas aos hiperplanos na Figura 30 são chamadas de vetores de suporte, pois uma variação na posição dessas amostras significaria uma variação na posição do hiperplano, já que ele está posicionado de modo a estar o mais distante possível

[2] Se os dados existem em P dimensões, um hiperplano é um subespaço de dimensão $P - 1$: para dados bidimensionais, um hiperplano é uma reta; para dados tridimensionais, um hiperplano é um plano; para dados de dimensões maiores a visualização dos hiperplanos é mais complexa, mas a ideia se mantém (JAMES et al., 2013, Seção 9.1.1).

das amostras de treinamento. O método recebe o nome de SVM devido a esses vetores, que são o suporte do hiperplano classificador (JAMES et al., 2013, Cap. 9).

Figura 30 – Hiperplanos de separação e observações bidimensionais de duas classes, a branca e a preta. H1 é um hiperplano que não é capaz de separar as classes. H2 é um hiperplano que é capaz de separar as duas classes, mas com uma margem pequena. H3 é um hiperplano de margem máxima, que separa as duas classes estando o mais distante possível das observações.



Fonte: (WEINBERG, 2012)

Existem casos em que não existe um hiperplano que separe completamente as classes. O SVM funciona, nesses casos, permitindo que algumas amostras ultrapassem as margens, ou seja, permitindo que algumas amostras sejam classificadas incorretamente. Isso faz com que o SVM seja mais robusto, sacrificando a classificação de algumas amostras para classificar com maior precisão a maioria das amostras (JAMES et al., 2013, Seção 9.2.2). A flexibilidade dessas margens é definida pelo parâmetro C, que limita o número e a severidade das violações às margens que podem ser permitidas. O parâmetro C controla o viés e a variância do SVM: um valor baixo de C resulta em margens estreitas raramente violadas e num classificador que se molda muito ao dados de treinamento, com baixo viés e alta variância; um valor alto de C resulta em margens mais largas e num classificador menos flexível, com alto viés e baixa variância.

Para aumentar a flexibilidade da classificação, o método da SVM se utiliza de *kernels*, que criam mapeamentos não-lineares a partir dos dados, o que aumenta a dimensionalidade do problema, mas possibilita a definição de limites lineares entre classes que, a princípio, não eram linearmente separáveis. Para classificar cada nova observação, a SVM avalia a

função

$$f(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle, \tag{5.9}$$

que é a soma de um parâmetro β_0 com o somatório dos produtos internos entre a nova observação x e cada uma das observações de treinamento x_i que são vetores de suporte. Esses produtos internos são multiplicados pelos parâmetros $\alpha_1, \ldots, \alpha_n$ (JAMES et al., 2013, Seção 9.3). Os *kernels* são generalizações da operação de produto interno na forma $K(x_i, x_{i'})$ que quantificam a similaridade entre duas observações (JAMES et al., 2013, Seção 9.3). O *kernel* mais simples é o linear, que quantifica a similaridade entre observações por meio da correlação de Pearson

$$K(x_i, x_{i'}) = \sum_{j=1}^{p} x_{ij} * x_{i'j}. \tag{5.10}$$

Outra opção, mais flexível é o *kernel* radial, dado pela seguinte equação

$$K(x_i, x_{i'}) = \exp\left(-\gamma \sum_{j=1}^{p} (x_{ij} * x_{i'j})^2\right), \tag{5.11}$$

sendo γ uma constante positiva. O *kernel* radial dá um peso maior para as observações de treinamento x_i que estão próximas da observação de teste x, agindo deste modo de forma local. Observações de treinamento muito distantes de x não exercem grande influência em sua classificação. Isso se verifica pela formulação matemática do *kernel* radial na Equação 5.11: se a distância $\sum_{j=1}^{p} (x_{ij} * x_{i'j})^2$ for grande, o valor de $K(x_i, x_{i'})$ será muito pequeno, devido à multiplicação pelo oposto da constante γ, seguido da operação exp().

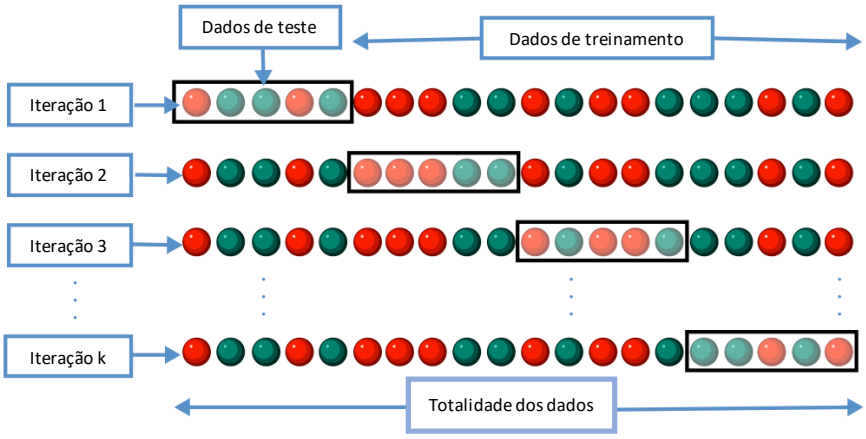
Até agora o SVM foi descrito com base em problemas de classificação com $k = 2$ classes. Contudo, muitos problemas, inclusive o problema dessa pesquisa, lidam com um número de classes $k > 2$. Há algumas abordagens criadas para possibilitar o uso de SVMs nesses casos, e as duas mais comuns são 1) Classificação um-contra-um, e 2) Classificação um-contra-todos (JAMES et al., 2013, Seção 9.4). A abordagem um-contra-um cria, para cada observação de teste x, $\binom{k}{2}$ SVMs, cada um comparando um par de classes. A observação de teste é, por fim, classificada na classe à qual ela foi mais frequentemente classificada nos $\binom{k}{2}$ SVMs. A abordagem um-contra-todos cria, para cada observação de teste x, k SVMs comparando umas das k classes com as $k - 1$ classes restantes. A observação de teste é, por fim, classificada na classe que obteve maior confiança quando foi comparada contra todas as outras.

5.4 Validação cruzada

Como discutido na Seção 5.1, a acurácia de teste de um classificador é uma medida de desempenho mais rigorosa do que a acurácia de treinamento, visto que os dados de teste são desconhecidos. Mas como dados de teste nem sempre estão disponíveis e ainda há a necessidade de medir o desempenho dos classificadores de forma rigorosa, são utilizados os métodos de reamostragem. Esse métodos envolvem repetidos treinamentos e testes de um modelo com diferentes subconjuntos dos dados de treinamento e sua utilidade está no fato de serem capazes de estimar a acurácia de teste dos classificadores. Essa performance é mais significativa do que a medida pelo erro de treinamento, que nada mais é do que o método classificando dados que ele já conhece, ou seja, que foram utilizados em seu treinamento, pois reflete a capacidade do método de lidar com eventuais dados novos.

Um dos métodos de reamostragem mais utilizado é a validação cruzada em K partes (KFCV, do inglês *K-fold cross validation*). Essa abordagem consiste em dividir de forma aleatória o conjunto total de observações em K partes de tamanhos iguais ou praticamente iguais, treinar o modelo com $K-1$ dessas partes e então validar o modelo com a parte restante, não utilizada no treinamento. Esse procedimento é repetido K vezes de modo que cada uma das partes é utilizada para validação uma única vez. O procedimento da KFCV é ilustrado na Figura 31.

Figura 31 – Procedimento realizado na validação cruzada em K partes com $K = k$. Em cada uma das k iterações, uma parte dos dados é utilizada para teste do modelo e o restante é utilizado para treinamento do modelo. O conjunto de dados ilustrado possui duas classes: a verde e a vermelha.



Fonte: Adaptado de (GUFOSOWA, 2019).

A estimativa da KFCV para a acurácia de teste é a média das acurácias obtidas em cada uma das K validações do modelo. Empiricamente, os valores de $K = 5$ e $K = 10$

para a KFCV produzem resultados que não sofrem nem de alto viés nem de alta variância e são geralmente utilizados (JAMES et al., 2013, Seção 5.1).

6 Resultados e Discussão

Este capítulo apresentará os resultados relativos aos objetivos do trabalho com a base de dados descrita no Capítulo 4. Além disso, apresentará discussões sobre os resultados, com o intuito de responder os objetivos específicos dessa pesquisa. O capítulo inicia com a descrição de como os métodos de classificação estatística descritos no Capítulo 5 foram aplicados à base de dados. As três seções seguintes deste capítulo abordarão, para as três línguas analisadas, resultados relativos à contribuição da informação visual para a percepção de tons lexicais, partindo de uma análise mais ampla (comparando tipos de sinais, na Seção 6.2) até análises mais específicas (comparando tons lexicais na Seção 6.3 e movimentos específicos da face e da cabeça na Seção 6.4). A última seção, por sua vez, apresenta uma discussão que busca sumarizar os resultados obtidos e sua relevância para o objetivo principal do trabalho.

6.1 Método de obtenção dos resultados

A abordagem utilizada neste trabalho para se obter resultados sobre a contribuição da componente visual da fala na percepção de tons lexicais foi o emprego de técnicas de classificação estatística. A questão é determinar quais partes da componente visual são mais importantes na classificação entre tons lexicais, e com nossa base de dados é possível medir essa capacidade de classificação dos quatro tipos de sinais descritos na Tabela 12. Além desses quatro sinais, serão utilizados como entrada para os classificadores mais dois sinais: um aleatório e um sinal que combina F0 com Movimento Total. O sinal Aleatório tem como objetivo ser um sinal completamente incapaz de prever sistematicamente tons lexicais: se o desempenho de um sinal for maior do que o do aleatório, esse sinal é capaz de prever sistematicamente tons lexicais em alguma medida. O sinal que combina F0 com Movimento Total tem o intuito de simular uma situação de comunicação real na qual informações acústicas e visuais estão simultaneamente disponíveis ao interlocutor. A Tabela 13 descreve esses sinais, sua composição e a nomenclatura utilizada para se referir a cada um deles a partir de agora.

Serão utilizados como classificadores dois métodos lineares (LDA e SVM linear) e dois métodos não-lineares (KNN e SVM radial), descritos na Seção 5.3. Métodos lineares são menos flexíveis e tendem a ter mais viés e menos variância do que métodos não-lineares, que são mais flexíveis.

A validação da acurácia obtida por cada método será realizada por meio do procedimento de validação cruzada em K partes (KFCV), descrito na Seção 5.4, com $K = 5$, pois esse é um valor que resulta num equilíbrio entre baixo viés e baixa variância

Tabela 13 – Sumarização dos sinais utilizados como entrada, com suas correspondentes descrições, dimensões e nomenclatura. Na coluna **Dimensão**, N se refere ao número de elocuções. O sinal F0 é a curva de F0 obtida a partir do sinal acústico. O sinal MT é o movimento total dos marcadores capturado pelo OPTOTRAK. Os sinais FC e CB são os sinas de movimento da face e da cabeça, respectivamente, obtidos a partir do procedimento de compensação do movimento da cabeça, descrito na Seção 4.2.2. O sinal F0+MT é a concatenação dos sinais F0 e MT. O sinal AL é um sinal aleatório uniforme.

Sinal	**Dimensão**	**Nomenclatura**
F0	$N \times 1$	F0
Movimento Total	$N \times 99$	MT
Face	$N \times 87$	FC
Cabeça	$N \times 6$	CB
$F0$+Movimento Total	$N \times (99 + 1)$	F0+MT
Aleatório	$N \times 1$	AL

(JAMES et al., 2013, Seção 5.1.4). Além disso, a validação cruzada em K partes será repetida 60 vezes para cada combinação de participante, tipo de sinal e método de classificação. Isso significa que, para cada participante (CNT1, CNT2, CNT3, MND1, MND2, MND3 e TAI1), cada tipo de sinal (F0 , MT, FC, CB, F0+MT e AL) servirá de entrada para cada um dos quatro métodos de classificação (KNN, LDA, SVM linear e SVM radial). O resultado de cada caso (combinação de participante/tipo de sinal/método) será uma distribuição de 60 valores, sendo cada um deles uma repetição de KFCV com $K = 5$, cujo resultado é a acurácia média obtida pelas 5 partes. Em cada uma das 60 repetições de KFCV, a divisão das 5 partes do conjunto de dados é diferente, resultando em acurácias diferentes, mas idealmente próximas entre si, para cada repetição.

Além de possibilitar o cálculo da dispersão das acurácias obtidas para cada caso, a repetição de 60 vezes da validação cruzada possibilita a comparação entre distribuições de acurácias de casos diferentes. Isso é útil, pois queremos saber se há diferenças nos resultados de uma língua para outra, ou de um método para outro, ou de um tipo de sinal para outro. Para isso, podem ser aplicados testes estatísticos para verificar se as acurácias obtidas em dois casos são significativamente diferentes ou não. Esses testes estatísticos serão realizados da seguinte maneira:

1. É verificada a normalidade de cada distribuição por meio do teste de Shapiro-Wilk ($p > 0,05$) (R CORE TEAM, 2019);

2. É verificada a homocedasticidade entre as distribuições sendo comparadas, ou seja, se as suas variâncias são iguais, pelo teste de Bartlett ($p > 0,05$) (R CORE TEAM, 2019);

3. Caso as distribuições comparadas sejam ambas normais e com variâncias iguais, é realizada uma análise de variância e um teste *post-hoc* par-a-par de Tukey para verificar a similaridade entre as distribuições ($p > 0,05$) (R CORE TEAM, 2019);

4. Caso alguma das distribuições não atenda ao critério de normalidade do teste de Shapiro-Wilk ou a homocedasticidade entre as distribuições não seja verificada, a análise de variância é realizada por meio do teste de Kruskal-Wallis (R CORE TEAM, 2019) e um teste *post-hoc* par-a-par de Dunn ($p > 0,05$) (OGLE; WHEELER; DINNO, 2020).

6.2 Contribuição dos diferentes tipos de sinais na classificação de tons lexicais

Esta seção apresentará resultados referentes aos seguintes objetivos específicos da pesquisa:

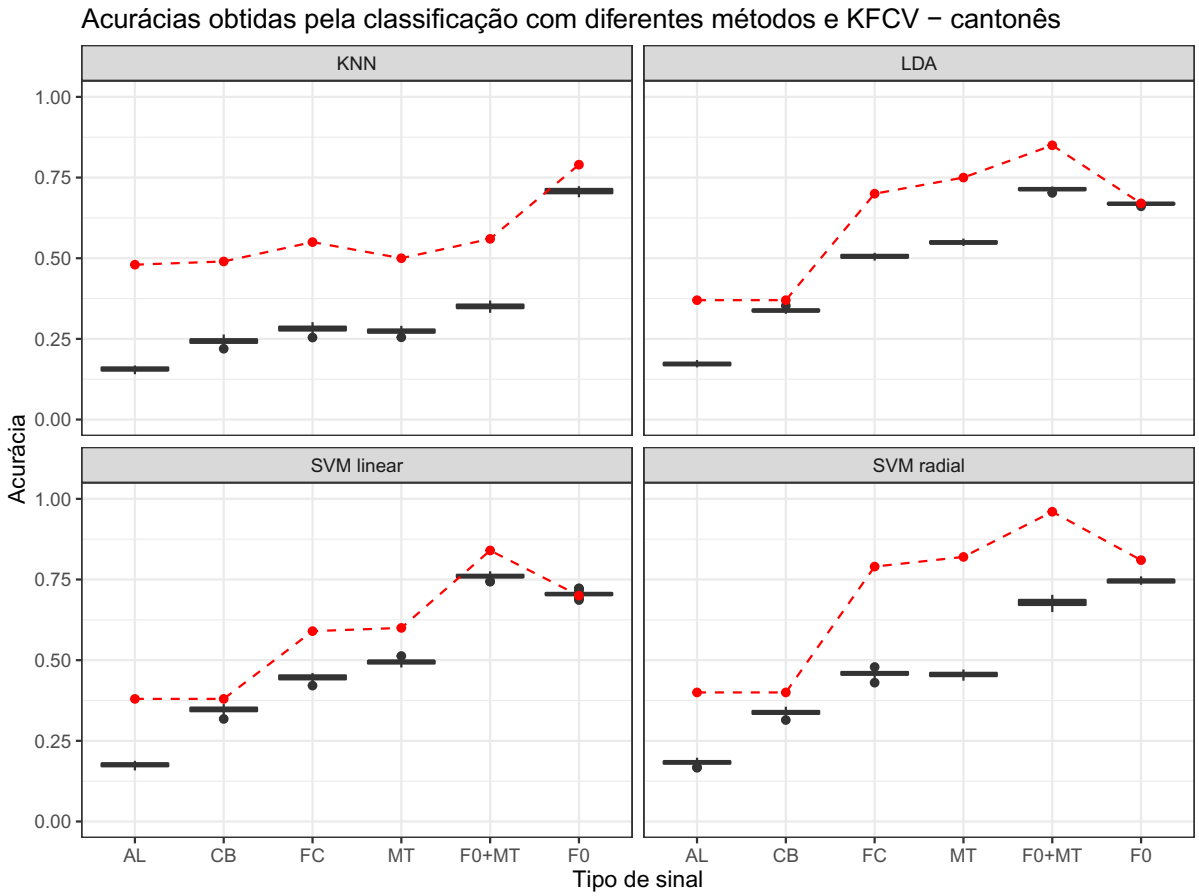
- determinar a capacidade da informação visual (sinais de movimento da face e da cabeça) de classificar tons lexicais e comparar essa capacidade com a do sinal acústico;
- determinar quais sinais de movimento (face ou cabeça) possuem maior capacidade de classificação de tons lexicais.

Para isso, são apresentadas as Figuras 32, 33 e 34, que trazem os resultados obtidos para cada um dos seis sinais de entrada em cada um dos quatro métodos de classificação para cantonês, mandarim e tailandês, respectivamente. Para o método KNN foi utilizado o valor de $K_{NN} \in \{5, 7, 9\}$ que obteve as maiores acurácias para cada sinal. A Tabela 14 traz os valores de K_{NN} utilizados para cada língua e para cada sinal.

Tabela 14 – Sumarização dos valores de K_{NN} utilizados para cada línguas e para cada sinal. Foi utilizado o valor de $K_{NN} \in \{5, 7, 9\}$ que obteve as maiores acurácias para cada caso.

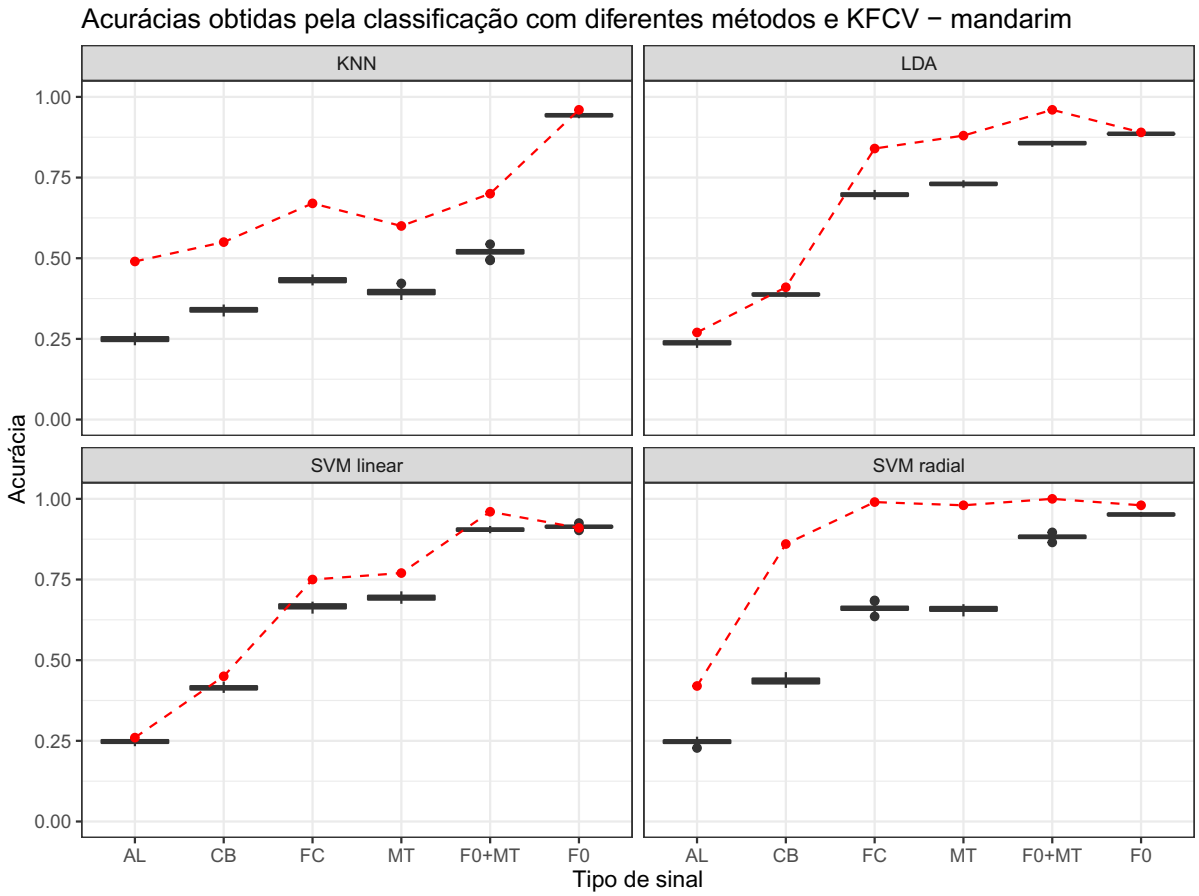
	AL	**CB**	**FC**	**MT**	**F0+MT**	**F0**
Cantonês	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$
Mandarim	$K_{NN} = 7$	$K_{NN} = 7$	$K_{NN} = 5$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 5$
Tailandês	$K_{NN} = 5$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 9$	$K_{NN} = 5$

Figura 32 – *Box plot* das acurácias obtidas pela classificação por meio dos métodos LDA, KNN, SVM linear e SVM radial para as três participantes falantes nativas de cantonês juntas. Cada painel traz as acurácias para cada um dos métodos de classificação. Em cada painel, a distribuição das 60 acurácias obtidas pelas 60 repetições de KFCV com $K = 5$ estão representadas em preto, e as acurácias de treinamento estão representadas em vermelho.



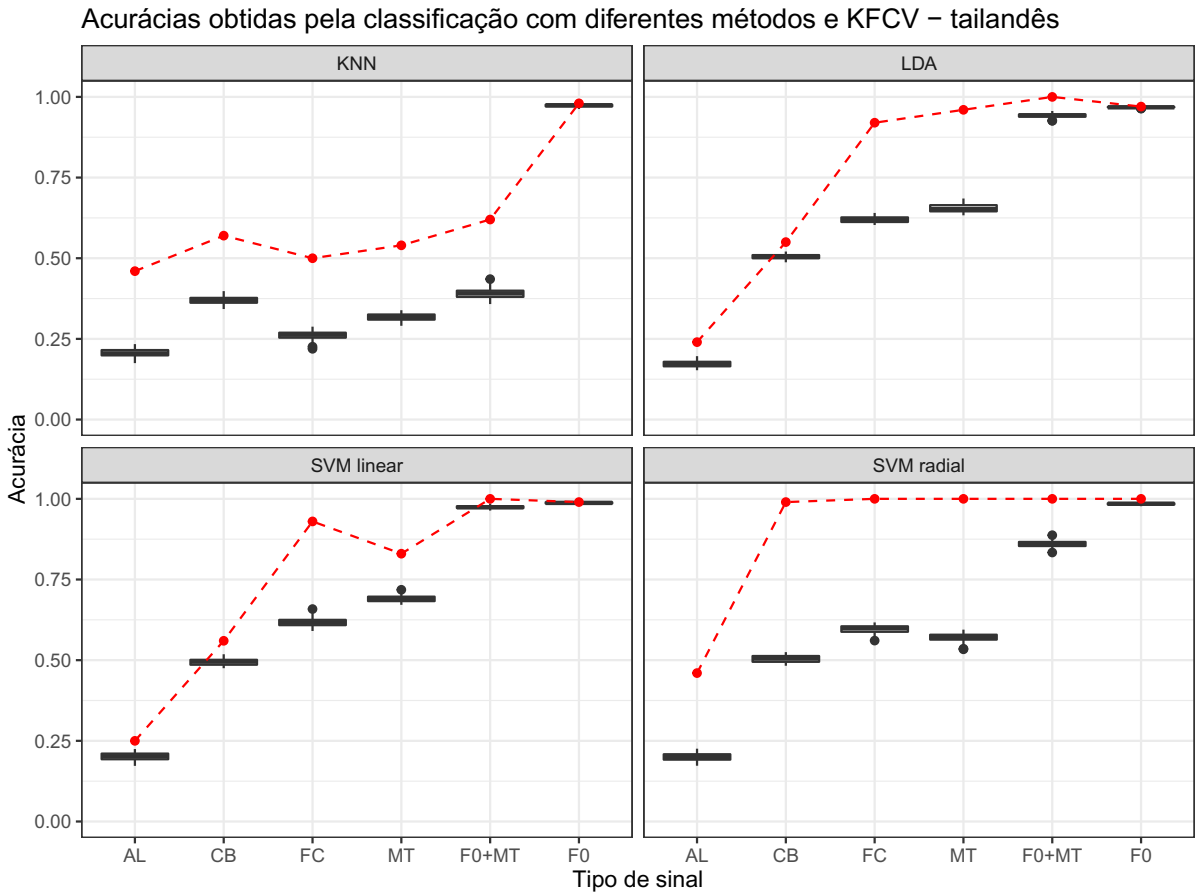
Fonte: o autor

Figura 33 – *Box plot* das acurácias obtidas pela classificação por meio dos métodos LDA, KNN, SVM linear e SVM radial para as três participantes falantes nativas de mandarim juntas. Cada painel traz as acurácias para cada um dos métodos de classificação. Em cada painel, a distribuição das 60 acurácias obtidas pelas 60 repetições de KFCV com $K = 5$ estão representadas em preto, e as acurácias de treinamento estão representadas em vermelho.



Fonte: o autor

Figura 34 – *Box plot* das acurácias obtidas pela classificação por meio dos métodos LDA, KNN, SVM linear e SVM radial para a participante RP. Cada painel traz as acurácias para cada um dos métodos de classificação. Em cada painel, a distribuição das 60 acurácias obtidas pelas 60 repetições de KFCV com $K = 5$ estão representadas em preto, e as acurácias de treinamento estão representadas em vermelho.



Fonte: o autor

Os próximos parágrafos tratarão dos resultados relativos às acurácias obtidas por meio de validação cruzada (KFCV), ilustradas em preto nas Figuras 32, 33 e 34, a não ser que o contrário esteja explícito. As Tabelas 15, 16 e 17 trazem os resultados dos testes estatísticos, descritos no fim da Seção 6.1, realizados para verificar diferenças significativas entre acurácias obtidas por sinais e métodos diferentes para cantonês, mandarim e tailandês, respectivamente.

Em relação à contribuição dos sinais de movimento da face e da cabeça para a classificação de tons lexicais, observou-se que todos os sinais puramente visuais (MT, FC e CB) obtiveram acurácias de classificação significativamente acima do sinal aleatório, o que significa que há componentes desses sinais que contribuem para a caracterização dos diferentes tons lexicais, tanto em cantonês, quanto em mandarim e tailandês. Por outro lado, nota-se que os sinais puramente visuais (MT, FC e CB) obtiveram acurácias significativamente menores do que as obtidas pelo sinal acústico, F0. Isso sugere que a componente acústica contribui mais para a classificação de tons lexicais do que a componente visual.

Entre os sinais com algum tipo de componente visual (F0+MT, MT, FC e CB), maiores acurácias foram obtidas em todos os métodos de classificação pelo sinal F0+MT, que é o sinal mais completo por possuir componentes acústicas e visuais. O desempenho dos outros sinais de movimento depende do tipo de método de classificação: 1) nos métodos lineares (LDA e SVM linear), para as três línguas, o sinal MT obteve maiores acurácias, seguido do sinal FC e do sinal CB, em ordem decrescente; 2) nos métodos não-lineares (KNN e SVM radial), para as três línguas, o sinal FC obteve sempre acurácias maiores ou estatisticamente iguais ao sinal MT, enquanto que o sinal CB obteve a menor acurácia entre os três (a exceção é em tailandês para o método KNN, em que a maior acurácia foi obtida pelo sinal CB, seguido de MT e, por fim, de FC). As menores acurácias foram as do sinal AL, que convergiram para o inverso do número de tons da língua ($1/6 = 16,67\%$ para cantonês, $1/4 = 25\%$ para mandarim e $1/5 = 20\%$ para tailandês). Isso somente não ocorreu para o erro de treinamento do sinal AL em alguns casos, devido provavelmente ao sobreajuste.

Comparando as acurácias obtidas pelos diferentes métodos com KFCV, percebe-se maiores acurácias dos métodos não-lineares (KNN e SVM radial) para o sinal F0 e maiores acurácias dos métodos lineares (LDA e SVM linear) para os sinais F0+MT, MT, FC e CB, aqueles com componentes visuais. Para alguns sinais, métodos diferentes produziram resultados estatisticamente iguais, conforme Tabelas 15, 16 e 17. Ressalta-se também o pior desempenho do método KNN, em comparação aos demais métodos, para os sinais de maior dimensionalidade (CB, FC, MT, F0+MT). Isso se deve ao fato de o método KNN basear sua classificação no cálculo da distância entre diferentes observações, e as distâncias calculadas entre elas, em um elevado número de dimensões, se tornarem muito

grandes e semelhantes umas às outras. Isso faz com que os dados se tornem mais esparsos e que as distâncias de uma observação a vizinhos de diferentes classes se tornem muito semelhantes, dificultando a classificação (JAMES et al., 2013, Seção 3.5).

Comparando as acurácias de treinamento (linhas vermelhas) com as acurácias obtidas pela KFCV (uma estimativa do erro de teste), percebe-se consistentemente que o método mais complexo (SVM radial) obteve maiores acurácias de treinamento do que todos os outros métodos. Contudo, o método SVM radial não obteve as maiores acurácias quando sujeito à validação cruzada, o que indica que o método sofre de sobreajuste, principalmente nos sinais de maior dimensão (F0+MT, MT, FC e CB), nos quais a diferença entre as acurácias de treinamento e de teste é maior.

Por fim, os resultados obtidos pelos métodos de classificação estatística utilizados nesse trabalho são comparáveis com resultados obtidos em experimentos de percepção. Em experimentos de percepção onde os participantes tinham que classificar os tons lexicais do cantonês sob três condições diferentes (AO, VO e AV), (BURNHAM; CIOCCA; STOKES, 2001; BURNHAM; LAU et al., 2001) observaram que tanto locutores nativos de cantonês quanto locutores nativos de tailandês e de inglês obtiveram acurácias na faixa de 80% a 85% na classificação de tons lexicais nas situações AO e AV, mas de apenas cerca de 20% em situações VO. As acurácias obtidas no presente trabalho são, em comparação às obtidas em (BURNHAM; CIOCCA; STOKES, 2001; BURNHAM; LAU et al., 2001), menores nas situações AO (sinal F0) e AV (sinal F0+MT), mas maiores nas situações VO (sinais CB, FC, MT). Em outro experimento, cujos estímulos consistiram de um vídeo da parte inferior da face de uma falante de mandarim e cujos participantes classificaram entre os diferentes tons lexicais em situações AO e AV, (MIXDORFF; HU; BURNHAM, 2005) obtiveram acurácias próximas de 100% para ambas as situações, resultados pouco acima dos obtidos no presente trabalho.

Tabela 15 – Sumarização dos testes estatísticos de diferença entre sinais e métodos para todas as participantes falantes nativas de cantonês juntas. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada linha dessa tabela, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE SINAIS			
Método	**Normalidade**	**Homocedasticidade**	**Sinais estatisticamente iguais**
KNN	✓	✓	-
LDA	✓	×	-
SVM L	×	✓	-
SVM R	✓	×	MT / FC
DIFERENÇA ESTATÍSTICA ENTRE MÉTODOS			
Sinal	**Normalidade**	**Homocedasticidade**	**Métodos estatisticamente iguais**
F0	✓	×	KNN / SVM linear
F0+MT	✓	×	-
MT	✓	×	-
FC	×	×	-
CB	×	×	LDA / SVM radial
AL	✓	✓	LDA / SVM linear

Tabela 16 – Sumarização dos testes estatísticos de diferença entre sinais e métodos para todas as participantes falantes nativas de mandarim juntas. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada linha dessa tabela, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE SINAIS			
Método	**Normalidade**	**Homocedasticidade**	**Sinais estatisticamente iguais**
KNN	✓	×	-
LDA	✓	×	-
SVM L	✓	×	-
SVM R	✓	×	MT / FC
DIFERENÇA ESTATÍSTICA ENTRE MÉTODOS			
Sinal	**Normalidade**	**Homocedasticidade**	**Métodos estatisticamente iguais**
F0	✓	×	-
F0+MT	✓	×	-
MT	✓	×	-
FC	✓	✓	-
CB	✓	×	-
AL	✓	✓	KNN / SVM linear / SVM radial

Tabela 17 – Sumarização dos testes estatísticos de diferença entre sinais e métodos para a participante falante nativa de tailandês juntas. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada linha dessa tabela, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE SINAIS			
Método	**Normalidade**	**Homocedasticidade**	**Sinais estatisticamente iguais**
KNN	×	×	-
LDA	×	×	-
SVM L	✓	×	-
SVM R	×	×	-
DIFERENÇA ESTATÍSTICA ENTRE MÉTODOS			
Sinal	**Normalidade**	**Homocedasticidade**	**Métodos estatisticamente iguais**
F0	×	×	-
F0+MT	✓	×	-
MT	×	✓	-
FC	×	×	LDA / SVM linear
CB	✓	×	LDA / SVM radial
AL	✓	✓	KNN / SVM lin. e SVM lin. / SVM rad.

6.3 Comparação entre tons lexicais

Esta seção apresentará resultados referentes ao seguinte objetivo específico da pesquisa: verificar se existem, em línguas específicas, tons lexicais que são mais fáceis do que outros de serem discriminados a partir da informação visual (sinais de movimento da face e da cabeça).

Para isso são apresentadas as tabelas 18, 19 e 20, que trazem as matrizes de confusão normalizadas obtidas para cada tipo de sinal com o método de classificação LDA, que é o método de classificação utilizado com a maior interpretabilidade (ver Seção 5.3.1) e que, além disso, apresentou melhor acurácia no geral.

Tabela 18 – Matrizes de confusão normalizadas resultantes da classificação por LDA com KFCV para todas as participantes de cantonês. Em cada matriz, a soma de cada coluna é igual a 1,00. Em parêntesis ao lado do tipo de sinal de cada matriz de confusão está sua acurácia total, considerando todos os tons.

		REFERÊNCIA												
		Aleatório (0,17)							**Cabeça** (0,34)					
		1	2	3	4	5	6		1	2	3	4	5	6
PREDIÇÃO	1	**0,23**	0,21	0,20	0,20	0,21	0,22	1	**0,56**	0,15	0,23	0,13	0,16	0,17
	2	0,15	**0,14**	0,15	0,13	0,15	0,14	2	0,08	**0,31**	0,11	0,10	0,22	0,15
	3	0,13	0,16	**0,12**	0,15	0,16	0,14	3	0,20	0,19	**0,31**	0,14	0,23	0,22
	4	0,27	0,26	0,30	**0,32**	0,27	0,28	4	0,04	0,09	0,09	**0,45**	0,09	0,15
	5	0,10	0,11	0,11	0,10	**0,09**	0,11	5	0,07	0,18	0,13	0,10	**0,21**	0,15
	6	0,12	0,12	0,12	0,10	0,12	**0,11**	6	0,05	0,08	0,13	0,08	0,09	**0,16**
		Face (0,51)							**Movimento Total** (0,55)					
		1	2	3	4	5	6		1	2	3	4	5	6
	1	**0,75**	0,06	0,11	0,02	0,05	0,04	1	**0,79**	0,05	0,08	0,02	0,04	0,04
	2	0,04	**0,55**	0,06	0,05	0,20	0,07	2	0,03	**0,59**	0,05	0,04	0,20	0,06
	3	0,14	0,09	**0,44**	0,08	0,21	0,25	3	0,12	0,09	**0,53**	0,08	0,21	0,25
	4	0,01	0,06	0,04	**0,65**	0,06	0,12	4	0,01	0,04	0,03	**0,70**	0,05	0,11
	5	0,03	0,18	0,16	0,08	**0,32**	0,20	5	0,03	0,18	0,14	0,06	**0,35**	0,19
	6	0,03	0,06	0,18	0,12	0,16	**0,32**	6	0,02	0,05	0,17	0,10	0,15	**0,35**
		F0 (0,67)							**F0 + Movimento Total** (0,71)					
		1	2	3	4	5	6		1	2	3	4	5	6
	1	**0,82**	0,00	0,06	0,00	0,08	0,00	1	**0,97**	0,01	0,01	0,00	0,01	0,01
	2	0,00	**0,73**	0,00	0,02	0,31	0,01	2	0,00	**0,67**	0,01	0,01	0,24	0,03
	3	0,17	0,00	**0,67**	0,01	0,19	0,19	3	0,02	0,04	**0,77**	0,01	0,20	0,17
	4	0,00	0,01	0,01	**0,82**	0,01	0,04	4	0,00	0,02	0,01	**0,88**	0,01	0,06
	5	0,01	0,17	0,06	0,00	**0,26**	0,04	5	0,00	0,21	0,09	0,01	**0,43**	0,14
	6	0,00	0,09	0,20	0,15	0,15	**0,72**	6	0,01	0,05	0,11	0,09	0,11	**0,59**

Tabela 19 – Matrizes de confusão normalizadas resultantes da classificação por LDA com KFCV para todas as participantes de mandarim. Em cada matriz, a soma de cada coluna é igual a 1,00. Em parêntesis ao lado do tipo de sinal de cada matriz de confusão está sua acurácia total, considerando todos os tons.

		REFERÊNCIA								
		Aleatório (0,24)					**Cabeça** (0,39)			
		1	2	3	4		1	2	3	4
PREDIÇÃO	1	**0,29**	0,29	0,30	0,30	1	**0,48**	0,33	0,19	0,27
	2	0,21	**0,20**	0,24	0,24	2	0,17	**0,27**	0,16	0,18
	3	0,36	0,37	**0,33**	0,34	3	0,20	0,27	**0,52**	0,27
	4	0,14	0,14	0,13	**0,12**	4	0,15	0,13	0,13	**0,28**
		Face (0,70)					**Movimento Total** (0,73)			
		1	2	3	4		1	2	3	4
	1	**0,69**	0,15	0,05	0,14	1	**0,73**	0,15	0,05	0,12
	2	0,16	**0,65**	0,16	0,06	2	0,13	**0,67**	0,15	0,05
	3	0,05	0,16	**0,73**	0,07	3	0,05	0,14	**0,74**	0,06
	4	0,10	0,04	0,06	**0,73**	4	0,09	0,04	0,06	**0,77**
		F0 (0,89)					**F0 + Movimento Total** (0,86)			
		1	2	3	4		1	2	3	4
	1	**0,95**	0,01	0,05	0,09	1	**0,87**	0,05	0,03	0,08
	2	0,04	**0,95**	0,18	0,01	2	0,06	**0,85**	0,09	0,04
	3	0,00	0,04	**0,75**	0,00	3	0,02	0,07	**0,84**	0,02
	4	0,01	0,00	0,02	**0,90**	4	0,05	0,03	0,04	**0,86**

Tabela 20 – Matrizes de confusão normalizadas resultantes da classificação por LDA com KFCV para a participante de tailandês. Em cada matriz, a soma de cada coluna é igual a 1,00. Em parêntesis ao lado do tipo de sinal de cada matriz de confusão está sua acurácia total, considerando todos os tons.

		REFERÊNCIA										
		Aleatório (0,17)						**Cabeça** (0,50)					
		1	2	3	4	5		1	2	3	4	5
PREDIÇÃO	1	**0,27**	0,29	0,33	0,28	0,28	1	**0,39**	0,12	0,19	0,18	0,07
	2	0,19	**0,17**	0,21	0,24	0,22	2	0,18	**0,49**	0,06	0,18	0,21
	3	0,19	0,14	**0,10**	0,15	0,16	3	0,16	0,06	**0,60**	0,05	0,01
	4	0,20	0,22	0,18	**0,16**	0,20	4	0,20	0,15	0,11	**0,47**	0,15
	5	0,15	0,18	0,18	0,17	**0,14**	5	0,07	0,18	0,04	0,12	**0,56**
		Face (0,62)						**Movimento Total** (0,65)				
		1	2	3	4	5		1	2	3	4	5
	1	**0,50**	0,17	0,17	0,12	0,08	1	**0,56**	0,14	0,18	0,14	0,06
	2	0,17	**0,61**	0,08	0,06	0,10	2	0,13	**0,70**	0,08	0,04	0,09
	3	0,15	0,06	**0,63**	0,08	0,02	3	0,16	0,05	**0,62**	0,06	0,02
	4	0,13	0,07	0,09	**0,64**	0,13	4	0,10	0,04	0,09	**0,68**	0,13
	5	0,05	0,09	0,03	0,10	**0,67**	5	0,05	0,07	0,03	0,08	**0,70**
		F0 (0,97)						**F0 + Movimento Total** (0,94)				
		1	2	3	4	5		1	2	3	4	5
	1	**1,00**	0,03	0,04	0,02	0,00	1	**0,92**	0,04	0,03	0,07	0,00
	2	0,00	**0,97**	0,00	0,01	0,02	2	0,03	**0,93**	0,00	0,01	0,02
	3	0,00	0,00	**0,95**	0,00	0,00	3	0,01	0,00	**0,97**	0,01	0,00
	4	0,00	0,00	0,01	**0,94**	0,01	4	0,04	0,01	0,00	**0,90**	0,01
	5	0,00	0,00	0,00	0,03	**0,97**	5	0,00	0,02	0,00	0,01	**0,97**

As tabelas 21, 22 e 23 trazem os resultados dos testes estatísticos, descritos no fim da Seção 6.1, realizados para verificar diferenças estatisticamente significativas entre acurácias obtidas por tons diferentes em cantonês, mandarim e tailandês, respectivamente.

Tabela 21 – Sumarização dos testes estatísticos de diferença entre tons lexicais para todas as participantes falantes nativas de cantonês juntas e para o método de classificação LDA. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada tom lexical em cada tipo de sinal, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE TONS LEXICAIS	
Tipo de sinal	**Tons estatisticamente iguais**
F0	1 / 4
F0+MT	-
MT	-
FC	-
CB	2 / 3
AL	2 / 3

Tabela 22 – Sumarização dos testes estatísticos de diferença entre tons lexicais para todas as participantes falantes nativas de mandarim juntas e para o método de classificação LDA. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada tom lexical em cada tipo de sinal, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE TONS LEXICAIS	
Tipo de sinal	**Tons estatisticamente iguais**
F0	-
F0+MT	1 / 4
MT	-
FC	3 / 4
CB	-
AL	-

Tabela 23 – Sumarização dos testes estatísticos de diferença entre tons lexicais para a participante falante nativa de tailandês e para o método de classificação LDA. A normalidade de cada distribuição foi aferida pelo teste de Shapiro-Wilk ($p > 0,05$) e a homocedasticidade de cada conjunto de distribuições foi aferida pelo teste de Bartlett ($p > 0,05$). Para cada tom lexical em cada tipo de sinal, se há normalidade e homocedasticidade, foi realizada análise de variância e posterior teste *posthoc* par-a-par pelo teste de Tukey ($p > 0,05$) para verificar similaridade entre distribuições. Caso não haja normalidade e homocedasticidade, a análise de variância foi realizada pelo teste de Kruskal-Wallis com posterior teste *posthoc* de Dunn.

DIFERENÇA ESTATÍSTICA ENTRE TONS LEXICAIS	
Tipo de sinal	**Tons estatisticamente iguais**
F0	-
F0+MT	1 / 4
MT	2 / 4, 2 / 5, 4 / 5
FC	2 / 3
CB	2 / 4
AL	-

Em cantonês, os sinais com algum tipo de componente visual (F0+MT, MT, FC e CB) apresentaram maiores acurácias na classificação dos tons 1 e 4, com o tom 1 obtendo a maior acurácia de todas. Por outro lado, para os mesmos sinais, as menores acurácias foram obtidas na classificação dos tons 5 e 6, com os tons 2 e 3 obtendo acurácias intermediárias. O sinal F0, puramente acústico, obteve, por sua vez, a menor acurácia na classificação do tom 5, classificado mais vezes como tom 2 do que como tom 5.

Em mandarim, os sinais com algum tipo de componente visual (F0+MT, MT, FC e CB) obtiveram resultados que parecem depender da presença do sinal de movimento da face. Quando o sinal de movimento da face está ausente (sinal CB), a acurácia média é baixa (39%) e os tons 2 e 4 são classificados com acurácias significativamente menores do que os tons 1 e 3. Por outro lado, quando o sinal de movimento da face está presente (sinais F0+MT, MT e FC) as acurácias médias aumentam (70%) e a variação da acurácias entre tons diminui (8%). Isso sugere que o movimento da face contribui de forma mais significativa do que o sinal de movimento da cabeça para a classificação de tons lexicais em mandarim. O sinal F0, puramente acústico, obteve a maior acurácia média entre todos os sinais, mas classificou o tom 3 significativamente pior do que os outros tons, o que pode se dever à maior complexidade de se estimar F0 para este tom específico, que possui uma mudança de direção ao longo de sua produção: começa descendente e termina crescente.

Em tailandês, os sinais MT, FC e CB, com apenas componentes visuais, apresentaram a menor acurácia para o tom 1. O sinal CB obteve maior acurácia para o tom 3, enquanto que os sinais MT e FC obtiveram maiores acurácias para os tons 2, 4, 5 e 4, 5, respectivamente. Por outro lado, o sinais com algum tipo de componente acústica (F0 e F0+MT) obtiveram maiores acurácias médias e pouca diferença entre os diferentes tons.

6.4 Análise de componentes visuais específicos

Esta seção apresentará resultados referentes ao seguinte objetivo específico da pesquisa: determinar quais componentes dos sinais de movimento da face e da cabeça são mais relevantes para a classificação de tons lexicais.

Para isso, podemos analisar a composição de cada uma das dimensões (LDs) em que a LDA projetou os dados. Quanto maior a contribuição de sinal específico, por exemplo, o movimento no eixo X da sobrancelha, na combinação linear que define cada LD, maior sua influência na separabilidade entre tons lexicais. As tabelas 24, 25 e 26 trazem os componentes mais influentes para as LDs definidas pela LDA para os sinais MT em cantonês, mandarim e tailandês, respectivamente.

É importante ressaltar que, ao contrário de outras técnicas de redução de dimensionalidade como a Análise de Componentes Principais (PCA) (JAMES et al., 2013, Seção 10.2), que projeta os dados originais em dimensões ortogonais entre si, a LDA projeta os dados em dimensões que não são necessariamente ortogonais entre si. Ou seja, LDs diferentes podem compartilhar as mesmas componentes. Além disso, da mesma forma que para a PCA, a LDA produz LDs ordenadas pela variância do conjunto de dados total explicada, o que faz com que a primeira LD (LD1) represente maior variância do que a LD2 e assim por diante. Além disso, há outra diferença entre a redução de dimensionalidade por meio da PCA e por meio da LDA: enquanto a PCA determina dimensões que maximizem a variância dos dados, a LDA determina dimensões que maximizem a separabilidade entre classes. Deste modo, a LDA pode ser mais apropriada para problemas de classificação do que a PCA, pois sua formulação matemática privilegia a separação entre as classes.

Tabela 24 – Componentes mais influentes das 5 LDs definidas pela LDA para os sinais MT de todas as participantes de cantonês juntas. Aqui as componentes mais influentes são aquelas que colaboram com, no mínimo, 75% do seu valor original na combinação linear de cada LD.

LD	**Sinais mais influentes**
LD1	Corpo rígido (y) / Sobrancelha direita (x) / Laringe (x)
LD2	Corpo rígido (y) / Lábio inferior (y)
LD3	Corpo rígido (y)
LD4	Corpo rígido (y)
LD5	Corpo rígido (y)/ Bochechas (y)/ Ponta do nariz (x, z)

Para cantonês, há presença do sinal do corpo rígido em todas as LDs, o que representa a influência do movimento da cabeça na separabilidade entre tons. Há também a influência de sinais ligados diretamente à fonação, como o movimento da laringe, e à articulação necessária para a fala, como o movimento do lábio inferior e das bochechas. O movimento da sobrancelha também se faz presente, assim como o movimento da ponta

Tabela 25 – Componentes mais influentes das 3 LDs definidas pela LDA para os sinais MT de todas as participantes de mandarim juntas. Aqui as componentes mais influentes são aquelas que colaboram com, no mínimo, 75% do seu valor original na combinação linear de cada LD.

LD	**Sinais mais influentes**
LD1	Corpo rígido (y) / Sobrancelhas (x, y)
LD2	Sobrancelha esquerda (x)
LD3	Sobrancelhas (y)

Tabela 26 – Componentes mais influentes das 4 LDs definidas pela LDA para os sinais MT da participante de tailandês. Aqui as componentes mais influentes são aquelas que colaboram com, no mínimo, 75% do seu valor original na combinação linear de cada LD.

LD	**Sinais mais influentes**
LD1	Lábio inferior (y)
LD2	Lábio inferior (y)
LD3	Ponta do nariz (y) / Laringe (y)
LD4	Mandíbula (y)

do nariz, que pode aqui ser interpretado como um movimento totalmente dependente do movimento da cabeça.

Para mandarim, o movimento da cabeça está presente na primeira LD, acompanhado pelo movimento das sobrancelhas em todas as outras LDs.

Para tailandês, o movimento da cabeça está presente por meio do movimento da ponta do nariz, que não se move independentemente da cabeça. Os outros sinais influentes são relacionados diretamente à produção da fala, como o movimento do lábio inferior, da laringe e da mandíbula.

6.5 Discussão

Primeiramente, cabe ressaltar que a motivação para o uso de classificadores (que, até onde sabemos, é um método novo) neste trabalho se baseia naquilo que entendemos que os humanos fazem numa situação de comunicação por meio fala: eles precisam decidir qual o tom lexical que está contido naquela informação, acústica e/ou visual. No entanto, não podemos assumir que o processo relativamente simples realizado pelos classificadores propostos no trabalho correspondem aos processos certamente muito mais complexos associados à percepção humana.

Os resultados deste trabalho evidenciam uma limitação do processo, que não nos permite traçar conclusões gerais para nenhuma das línguas estudadas. Dados medidos em experimentos de produção de fala trazem, do ponto de vista estatístico, componentes inerentes 1) à língua, 2) ao locutor e 3) ao conteúdo linguístico sendo falado. Para podermos tirar conclusões estatisticamente válidas a respeito da relação entre o desempenho dos classificadores e qualquer um destes três fatores, é necessária uma amostragem grande dos outros dois fatores. Isso é especialmente crítico neste trabalho para o tailandês, pois nossa base de dados é composta por apenas uma locutora da língua, de tal forma que estatisticamente não se pode afirmar se os resultados são inerentes à língua ou ao locutor.

O movimento da cabeça possui uma forte correlação com F0(YEHIA; KURATATE; VATIKIOTIS-BATESON, 2002), com correlatos mais específicos sendo obtidos com unidades prosódicas mais longas que o tom lexical, como por exemplo o *pitch accent* (KRAHMER; SWERTS, 2007). Os resultados obtidos neste trabalho indicam que o movimento da cabeça (sinal CB) possui maior capacidade de classificar tons de nível alto (tons caracterizados pela produção constante de frequência aguda) em todas as línguas, de acordo com as tabelas 18, 19 e 20. Tanto em cantonês quanto em tailandês, os tons lexicais melhor classificados pelo movimento da cabeça foram os tons de nível alto (tom 1 em cantonês e tom 3 em tailandês). Em mandarim o tom 1, de nível alto, obteve a segunda maior acurácia, logo atrás do tom 3, que é um tom com uma mudança de direção ao longo de sua duração, tornando-o mais longo (BLICHER; DIEHL; COHEN, 1990) e diferenciando-o dos outros 3 tons do mandarim (HOOLE; HU, 2004). À luz da correlação entre F0 e movimento da cabeça, nossos resultados sugerem relação entre movimento da cabeça e tons lexicais de nível alto, que são produções contínuas de F0 em nível alto.

Além disso, nossos resultados para o cantonês estão em concordância com os apresentados por (BARRY; BLAMEY, 2004), que definem os tons 1 e 4 como os dois tons mais diferenciados dos demais tons, que se sobrepõem de forma mais significativa. Para todos os sinais, as maiores acurácias de classificação foram obtidas para os tons 1 e 4, sugerindo que eles se diferenciam de forma mais clara dos demais.

Outro resultado foram as diferenças obtidas entre as contribuições dos sinais FC

e CB. Para todas as línguas e para todos os métodos de classificação (com exceção do KNN), foram obtidas maiores acurácias com o sinal FC do que com o sinal CB. Isso sugere que a contribuição do movimento da face é maior do que a contribuição do movimento da cabeça na classificação de tons lexicais para as três línguas estudadas. Apesar disso, o movimento da cabeça ainda contribui para a classificação de tons lexicais, como mostram as tabelas 24 e 25, na Seção 6.4.

Por outro lado, as acurácias obtidas quando os sinais de movimento da face e da cabeça estão simultaneamente presentes (sinais F0+MT e MT) nunca são menores do que as acurácias obtidas com cada um desses sinais separadamente quando o método de classificação é linear. Isso sugere que as contribuições dos movimentos da face e da cabeça para a classificação de tons lexicais são diferentes entre si e, de alguma forma, complementares.

É possível também comparar as acurácias obtidas com os sinais de três tipos: aqueles com componentes apenas visuais (MT, FC e CB), aqueles com componentes apenas acústicas (F0) e aqueles com componentes audiovisuais (F0+MT). Levando em conta a percepção humana, espera-se que quanto mais informações estiverem disponíveis, maior a inteligibilidade da fala. Nosso método, contudo, é uma forma diferente e mais simples de se classificar tons lexicais. Deste modo, algumas limitações do nosso método são esperadas, sendo as principais as seguintes:

- Não se espera uma classificação perfeita (100% de acurácia) do nosso classificador como se esperaria de um falante nativo em condições ideais;
- Não se espera que nosso método tenha o mesmo comportamento com sinais de dimensionalidade alta que tem com sinais de dimensionalidade baixa, devido à maldição da dimensionalidade (BISHOP, 2006, Seção 1.4) (JAMES et al., 2013, Seção 6.4).

Com base nessas limitações, dois pontos podem ser discutidos. O primeiro é que o sinal F0+MT nem sempre obteve a maior das acurácias, mesmo sendo o sinal com mais informação disponível. Isso se deve a uma limitação dos métodos de classificação baseados em aprendizado estatístico: o sobreajuste, que ocorre quando o método se torna específico demais, obtendo alta acurácia no conjunto de treinamento e baixa acurácia em conjuntos de teste. Observou-se consistentemente em todas as línguas e métodos (à exceção do KNN) maior acurácia de treinamento para o sinal F0+MT do que para todos os outros. Contudo, os erros de teste, estimados pela validação cruzada em K partes (KFCV), nem sempre refletiram a mesma situação: apenas em cantonês e nos métodos lineares de classificação (LDA e SVM linear) é que o sinal F0+MT obteve a maior acurácia de teste de todas. Em todos os outros casos, a maior acurácia foi obtida pelo sinal de F0, o sinal com menor dimensionalidade entre todos, de acordo com a Tabela 12.

Um segundo ponto é o fato de a acurácia obtida pelos sinais com componentes apenas visuais (MT, FC e CB) ter sido sempre menor do que a acurácia obtida pelo sinal de F0, com componentes apenas acústicas. Isso vai de acordo com os resultados dos experimentos perceptivos realizados em (BURNHAM; LAU et al., 2001) e (BURNHAM; CIOCCA; STOKES, 2001), para o cantonês, e em (MIXDORFF; HU; BURNHAM, 2005), para o mandarim, em que a classificação dos tons lexicais em situações AO e AV foi significativamente maior do que a classificação em situações VO. Tons lexicais são tradicionalmente caracterizados em termos de F0 e, portanto, é de se esperar que F0 seja seu principal correlato físico (YIP, M., 2002). As acurácias altas obtidas pelo nosso método com o sinal F0 refletem sua capacidade de classificar, dado que a entrada seja apropriada. Contudo, nossos classificadores estatísticos não foram capazes de alcançar o desempenho obtido por falantes nativos em experimentos de percepção de tons lexicais em situações AO e AV (BURNHAM; LAU et al., 2001; BURNHAM; CIOCCA; STOKES, 2001; MIXDORFF; HU; BURNHAM, 2005). Isso provavelmente ocorre porque nem todas as informações utilizadas por participantes num experimento de percepção podem ser parametrizadas nos dados de entrada dos nossos classificadores (contexto, entoação, movimentos corporais etc.). Uma possível forma de melhorar o desempenho desses classificadores seria fornecendo informações adicionais, como por exemplo a duração das elocuções, que foi de certa maneira perdida neste trabalho com a realização da parametrização dos sinais por coeficientes polinomiais.

Além disso, nossos resultados sugerem que as componentes mais relevantes para a classificação de tons lexicais não são necessariamente aquelas associadas com uma variância mais elevada. Isso foi constatado através de uma análise que fizemos em que os dados foram submetidos à análise PCA antes de serem usados como entrada no classificador (LDA). Os resultados obtidos nesse caso (LDA precedida de PCA) foram piores do que aqueles obtidos apenas com LDA.

Uma outra questão é o fato de a LDA realizar a classificação num espaço de dimensão muito mais baixa que os outros métodos (KNN, SVM linear e SVM radial). Isso ocorre pois a LDA funciona em dois passos (ver Seção 5.3.1): o primeiro passo consiste numa redução de dimensionalidade dos dados (onde o espaço reduzido tem dimensão $N_c - 1$, sendo N_c o número de classes) e o segundo passo consiste na classificação propriamente dita neste espaço de dimensão reduzida. No entanto, essa redução de dimensionalidade não ocorre no caso dos outros classificadores, e portanto esta diferença de dimensionalidade provavelmente afetou os resultados. Uma forma de fazer uma comparação mais justa entre todos os métodos seria realizar redução de dimensionalidade por meio da LDA antes da classificação, de modo que todos os métodos lidem com os dados num mesmo número menor de dimensões.

Por fim, os resultados obtidos indicando que, para mandarim, o movimento das

sobrancelhas exerce grande influência na classificação dos tons lexicais estão em concordância com os resultados obtidos por (GARG et al., 2019), que associam as trajetórias descendentes e crescentes de F0 dos tons com os respectivos movimentos das sobrancelhas.

7 Conclusão

Resultados significativos no campo da fala multimodal em línguas tonais foram obtidos nas duas últimas décadas, principalmente relacionando a produção e a percepção de tons lexicais às componentes visuais da fala (BURNHAM; CIOCCA; STOKES, 2001; BURNHAM; LAU et al., 2001; MIXDORFF; HU; BURNHAM, 2005; CHEN; MASSARO, 2008; SMITH; BURNHAM, 2012; GARG et al., 2019; HAN et al., 2019). Este trabalho, além de utilizar uma metodologia diferente das utilizadas pelos estudos citados acima, que realizaram experimentos de percepção com uma série de participantes para aferir seus resultados, utilizou dados de produção de três línguas diferentes. Em relação à metodologia, a utilizada neste trabalho foi de classificação com base em técnicas de aprendizado estatístico, que a partir de seis tipos de sinais (F0, F0+MT, MT, FC, CB e AL) parametrizados por coeficientes polinomiais classificou cada palavra de acordo com seu tom lexical. Em relação ao número de participantes, foram utilizados dados gravados por três falantes nativas de cantonês, três falantes nativas de mandarim e uma falante nativa de tailandês, de modo que esses números são próximos aos utilizados nos outros estudos. Apesar de aplicar uma abordagem diferente dos testes de percepção, este trabalho obteve resultados em concordância com os obtidos pelos citados acima. Apesar disso, essa abordagem não substitui os experimentos de percepção, pois só trabalha com os dados da produção e os classifica (um procedimento que pode ser análogo à percepção).

Retornando à pergunta motivadora do trabalho, podemos agora sugerir uma resposta: **no contexto de uma abordagem quantitativa baseada na parametrização de informações audiovisuais da produção de tons lexicais, os movimentos da face e da cabeça contribuem para a percepção de tons lexicais em cantonês, mandarim e tailandês**. A classificação com base nas informações visuais da face e da cabeça apresentou, por um lado, acurácias sempre superiores do que a aleatória e, por outro, acurácias sempre inferiores àquelas obtidas pela classificação com base em F0 apenas. Nossos resultados, portanto, corroboram outros trabalhos da literatura, mostrando que o método proposto é capaz de utilizar a informação presente nos sinais de movimento da face e da cabeça para classificar tons lexicais. A pergunta que permanece, contudo, é: por quê a informação visual (sinais MT, FC e CB), que apresentou desempenho acima do aleatório, não foi capaz de melhorar o desempenho do F0 quando adicionada a ele (resultando no sinal F0+MT)? Em outras palavras, a questão é investigar a razão pela qual o desempenho do sinal F0+MT não foi sistematicamente melhor do que o desempenho do sinal F0. Esta é uma pergunta muito importante que não foi respondida neste trabalho e que, portanto, deveria ser abordada em trabalhos futuros.

Um ponto que não foi totalmente explorado neste trabalho diz respeito à deter-

minação das componentes dos sinais de movimento da face e da cabeça que são mais relevantes para a classificação de tons lexicais. Essa questão foi abordada aqui por meio da interpretação da LDA, que reduz a dimensionalidade dos dados de forma a maximizar a separabilidade entre classes. Abordagens mais apropriadas e específicas para essa tarefa, como o uso de técnicas de seleção de características, podem ser utilizadas em trabalhos futuros. Um exemplo de uso dessas técnicas seria através da combinação dos sinais F0, FC e CB numa única matriz de dados, de modo que um método de seleção de características pudesse sugerir respostas mais definitivas com relação a quais componentes são mais importantes para a classificação. Os resultados obtidos neste trabalho e sumarizados nas tabelas 24, 25 e 26 são, apesar disso, importantes, pois indicam possíveis candidatos para os movimentos da face e da cabeça mais relevantes na classificação de tons lexicais.

Uma possível validação perceptiva dos movimentos da face e da cabeça mais relevantes na classificação de tons lexicais (obtidos por meio do uso de técnicas de seleção de características) é pela realização de experimentos de percepção com estímulos baseados na síntese de avatares. Nesses experimentos, cada componente dos movimentos da face e da cabeça poderia ser manipulada individualmente, de forma que a relevância de cada uma possa ser analisada individualmente.

Outras sugestões de trabalhos futuros são a realização de experimentos de percepção com a mesma base de dados utilizada neste trabalho, a fim de verificar se os novos resultados obtidos estão em concordância com os apresentados aqui. Além disso, a adição de mais falantes e/ou línguas à base de dados por meio da realização de mais experimentos de coleta é uma forma de tornar os resultados mais gerais. Outra investigação, relativa ao método utilizado, que pode ser realizada no futuro é sobre a influência da duração da elocução nos coeficientes polinomiais, pois a duração pode ser um fator importante na caracterização de tons lexicais e isso permitiria que o papel disso na parametrização fosse melhor compreendido.

Por fim, uma vantagem do método apresentado é que ele pode funcionar com dados coletados de diferentes formas, uma vez que todos estejam parametrizados do mesmo modo no momento de serem classificados, o que o torna uma boa alternativa a trabalhos futuros na área de fala multimodal ou em áreas relacionadas.

Referências

BARRON, J L; FLEET, D J; BEAUCHEMIN, S S. Performance of Optical Flow Techniques. en, p. 60, 1994. Citado na p. 15.

BARRY, Johanna G.; BLAMEY, Peter J. The acoustic analysis of tone differentiation as a means for assessing tone production in speakers of Cantonese. **The Journal of the Acoustical Society of America**, v. 116, n. 3, p. 1739–1748, set. 2004. ISSN 0001-4966. DOI: 10.1121/1.1779272. Disponível em: <https://asa-scitation-org.ez27.periodicos.capes.gov.br/doi/abs/10.1121/1.1779272>. Acesso em: 6 jul. 2019. Citado na p. 98.

BISHOP, Christopher. **Pattern Recognition and Machine Learning**. New York: Springer-Verlag, 2006. (Information Science and Statistics). ISBN 978-0-387-31073-2. Disponível em: <https://www.springer.com/gp/book/9780387310732>. Acesso em: 18 jun. 2020. Citado nas pp. 76, 99.

BLICHER, Deborah L.; DIEHL, Randy L.; COHEN, Leslie B. Effects of syllable duration on the perception of the Mandarin Tone 2/Tone 3 distinction: evidence of auditory enhancement. en. **Journal of Phonetics**, v. 18, n. 1, p. 37–49, jan. 1990. ISSN 0095-4470. DOI: 10.1016/S0095-4470(19)30357-2. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095447019303572>. Acesso em: 20 jun. 2020. Citado na p. 98.

BOERSMA, Paul. Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound. en, p. 14, 1993. Citado nas pp. 54–58.

BOERSMA, Paul; WEENINK, David. **Praat: doing phonetics by computer**. 2020. Citado nas pp. 57, 58.

BROWMAN, Catherine P.; GOLDSTEIN, Louis M. Towards an Articulatory Phonology. **Phonology Yearbook**, v. 3, p. 219–252, 1986. ISSN 0265-8062. Disponível em: <https://www.jstor.org/stable/4615400>. Acesso em: 14 mar. 2020. Citado na p. 12.

BURNHAM, D.; LI, W. et al. **Visual correlates of Thai lexical tone production: Motion of the head, eyebrows, and larynx?** eng. 2019. Disponível em: <https://avsp2019.loria.fr/wp-content/uploads/2019/07/AVSP_2019_paper_16.pdf>. Acesso em: 2 mar. 2020. Citado na p. 19.

BURNHAM, Denis; CIOCCA, Valter; STOKES, Stephanie. Auditory-visual perception of lexical tone. In: INTERSPEECH. 2001. Citado nas pp. 16–19, 44, 45, 89, 100, 102.

BURNHAM, Denis; KASISOPA, Benjawan et al. Universality and language-specific experience in the perception of lexical tone and pitch. en. **Applied Psycholinguistics**, v. 36, n. 6, p. 1459–1491, nov. 2015. Publisher: Cambridge University Press. ISSN 0142-7164, 1469-1817. DOI: 10.1017/S0142716414000496. Disponível em: <https://www.cambridge.org/core/journals/applied-psycholinguistics/article/universality-and-languagespecific-experience-in-the-perception-of-lexical-tone-and-pitch/AF728C780F22599D7329C5654E57A29F>. Acesso em: 12 jul. 2020. Citado nas pp. 18, 19.

BURNHAM, Denis; LAU, Susanna et al. Visual Discrimination of Cantonese Tone by Tonal but Non-Cantonese Speakers, and by Non-Tonal Language Speakers. en, p. 6, 2001. Citado nas pp. 16–18, 44, 89, 100, 102.

CAVE, C. et al. About the relationship between eyebrow movements and Fo variations. In: PROCEEDING of Fourth International Conference on Spoken Language Processing. ICSLP '96. Out. 1996. v. 4, 2175–2178 vol.4. ISSN: null. Citado na p. 45.

CHAO, Yuen Ren. A system of tone letters. **Le Maître Phonétique**, v. 45, p. 24–27, 1930. Disponível em: <https://github.com/JacksonLLee/chao1930>. Citado na p. 40.

CHEN, Trevor H.; MASSARO, Dominic W. Seeing pitch: Visual information for lexical tones of Mandarin-Chinese. en. **The Journal of the Acoustical Society of America**, v. 123, n. 4, p. 2356–2366, abr. 2008. ISSN 0001-4966. DOI: 10.1121/1.2839004. Disponível em: <http://asa.scitation.org/doi/10.1121/1.2839004>. Acesso em: 28 jan. 2020. Citado nas pp. 16–18, 102.

CHRISTENSEN, Mads Græsbøll; JAKOBSSON, Andreas. **Multi-pitch estimation**. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool Publishers, 2009. OCLC: 1127127318. ISBN 978-1-59829-839-0. Disponível em: <http://VH7QX3XE2P.search.serialssolutions.com/?V=1.0&L=VH7QX3XE2P&S=JCs&C=TC0000329661&T=marc&tab=BOOKS>. Acesso em: 6 mai. 2020. Citado na p. 54.

COLEMAN, Thomas F.; LI, Yuying. An Interior Trust Region Approach for Nonlinean Minimization Subject to Bounds. Inglês. **SIAM Journal on Optimization**, v. 6, n. 2, p. 418–445, 1996. Citado na p. 61.

CRISTÓFARO-SILVA, Thaïs; YEHIA, Hani Camille. **Sonoridade em Artes, Saúde e Tecnologia**. 2009. Disponível em: <http://fonologia.org.>. Citado nas pp. 24, 25.

DANNER, Samantha Gordon; BARBOSA, Adriano Vilela; GOLDSTEIN, Louis. Quantitative analysis of multimodal speech data. en. **Journal of Phonetics**, v. 71, p. 268–283, nov. 2018. ISSN 00954470. DOI: 10.1016/j.wocn.2018.09.007. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0095447017302280>. Acesso em: 11 set. 2019. Citado nas pp. 15, 16.

DENES, P.B.; PINSON, E.N. **The Speech Chain: The Physics and Biology of Spoken Language, Second Edition**. Waveland Press, 2015. ISBN 978-1-4786-3107-1. Disponível em: <https://books.google.com.br/books?id=nYN2CgAAQBAJ>. Citado nas pp. 12, 13, 20, 29.

DIENER, Lorenz et al. An initial investigation into the real-time conversion of facial surface EMG signals to audible speech. eng. **Conference proceedings: ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference**, v. 2016, p. 888–891, 2016. ISSN 1557-170X. DOI: 10.1109/EMBC.2016.7590843. Citado na p. 14.

ERICKSON, Donna et al. Effect of Tone Height on Jaw and Tongue Articulation in Mandarin Chinese. en, p. 4, 2014. Citado na p. 44.

FANT, Gunnar. **Acoustic Theory of Speech Production, With Calculations based on X-Ray Studies of Russian Articulations**. Berlin, Boston: De Gruyter Mouton, 1960. ISBN 978-90-279-1600-6. DOI: 10.1515/9783110873429. Disponível em: <https://www.degruyter.com/view/product/140172>. Acesso em: 3 fev. 2020. Citado nas pp. 29, 30, 43, 54.

GARG, Saurabh et al. Computer-vision analysis reveals facial movements made during Mandarin tone production align with pitch trajectories. en. **Speech Communication**, v. 113, p. 47–62, out. 2019. ISSN 01676393. DOI: 10.1016/j.specom.2019.08.003. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0167639319301566>. Acesso em: 14 abr. 2020. Citado nas pp. 16–18, 101, 102.

GOLDSMITH, John A. **Autosegmental phonology**. 1976. Thesis – Massachusetts Institute of Technology. Disponível em: <http://www.ai.mit.edu/projects/dm/theses/goldsmith76.pdf>. Citado na p. 39.

GUFOSOWA. **K-fold cross validation EN**. 2019. Disponível em: <https://commons.wikimedia.org/wiki/File:K-fold_cross_validation_EN.svg>. Citado na p. 80.

HAN, Yueqiao et al. Relative Contribution of Auditory and Visual Information to Mandarin Chinese Tone Identification by Native and Tone-naïve Listeners. en. **Language and Speech**, p. 0023830919889995, dez. 2019. Publisher: SAGE Publications Ltd. ISSN 0023-8309. DOI: 10.1177/0023830919889995. Disponível em: <https://doi.org/10.1177/0023830919889995>. Acesso em: 31 mar. 2020. Citado nas pp. 17, 102.

HIROSE, Hajime. Investigating the Physiology of Laryngeal Structures. In: HARDCASTLE, William J.; LAVER, John; GIBBON, Fiona E. (Ed.). **The Handbook of Phonetic Sciences**. John Wiley & Sons, Ltd, 2010. P. 130–152. ISBN 978-1-4443-1725-1. DOI: 10.1002/9781444317251.ch4. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444317251.ch4>. Acesso em: 6 fev. 2020. Citado nas pp. 26, 28, 29.

HOOLE, Philip; HU, Fang. Tone-Vowel Interaction in Standard Chinese. en, p. 4, 2004. Citado nas pp. 44, 98.

HORN, Roger A.; JOHNSON, Charles R. **Matrix Analysis**. USA: Cambridge University Press, 1985. ISBN 0-521-30586-1. Citado na p. 63.

HU, Fang. Tonal Effect on Vowel Articulation in a Tone Language. In: citado na p. 44.

HYMAN, Larry M. How (not) to do phonological typology: the case of pitch-accent. en. **Language Sciences**, v. 31, n. 2-3, p. 213–238, mar. 2009. ISSN 03880001. DOI: 10.1016/j.langsci.2008.12.007. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0388000108000466>. Acesso em: 5 fev. 2020. Citado na p. 36.

_____. Tone Systems. In: HASPELMATH, M. et al. (Ed.). **Language Typology and Language Universals / Sprachtypologie und sprachliche Universalien / La typologie des langues et les universaux linguistiques. An International Handbook / Ein internationales Handbuch / Manuel international.** Boston, Berlin: De Gruyter Mouton, 2001. Citado na p. 37.

INTERNATIONAL PHONETIC ASSOCIATION. **IPA Chart**. 2015. Disponível em: <http://www.internationalphoneticassociation.org/content/ipa-chart>. Acesso em: 1 jul. 2020. Citado na p. 21.

JAMES, Gareth et al. **An Introduction to Statistical Learning**. New York, NY: Springer New York, 2013. v. 103. (Springer Texts in Statistics). ISBN 978-1-4614-7137-0 978-1-4614-7138-7. DOI: 10.1007/978-1-4614-7138-7. Disponível em: <http://link.springer.com/10.1007/978-1-4614-7138-7>. Acesso em: 29 abr. 2020. Citado nas pp. 64, 71, 72, 75–79, 81, 83, 89, 96, 99.

KARSTENS MEDIZINELEKTRONIK GMBH. **3D Electromagnetic Articulograph: Recording of Speech Movement**. Inglês. 2020. Disponível em: <https://www.articulograph.de/>. Acesso em: 27 jun. 2020. Citado na p. 14.

KINSLER, L.E. et al. **Fundamentals of Acoustics**. Wiley, 2000. ISBN 978-0-471-84789-2. Disponível em: <https://books.google.com.br/books?id=76IRAQAAIAAJ>. Citado nas pp. 29, 33.

KRAHMER, E. J.; SWERTS, M. G. J. The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. English. **Journal of Memory and Language**, v. 57, n. 3, p. 396–414, 2007. Publisher: ACADEMIC PRESS INC ELSEVIER SCIENCE. ISSN 0749-596X. Disponível em: <https://research.tilburguniversity.edu/en/publications/the-effects-of-visual-beats-on-prosodic-prominence-acoustic-analy>. Acesso em: 25 mai. 2020. Citado nas pp. 45, 98.

LADEFOGED, P.; JOHNSON, K. **A Course in Phonetics**. Cengage Learning, 2010. ISBN 978-1-4282-3126-9. Disponível em: <https://books.google.com.br/books?id=FjLc1XtqJUUC>. Citado nas pp. 21–25, 36.

LATIF, Nida et al. Movement Coordination during Conversation. en. Edição: Howard Nusbaum. **PLoS ONE**, v. 9, n. 8, e105036, ago. 2014. ISSN 1932-6203. DOI: 10.1371/journal.pone.0105036. Disponível em: <https://dx.plos.org/10.1371/journal.pone.0105036>. Acesso em: 28 jan. 2020. Citado na p. 16.

LIANG, Jie; HEUVEN, Vincent J. Chinese tone and intonation perceived by L1 and L2 listeners. In: LAHIRI, Aditi; GUSSENHOVEN, Carlos; RIAD, Tomas (Ed.). **Tones and Tunes: Experimental Studies in Word and Sentence Prosody**. Berlin, New York: Mouton de Gruyter, out. 2007. v. 2. P. 27–62. ISBN 978-3-11-019058-8 978-3-11-020757-6. DOI: 10.1515/9783110207576.1.27. Disponível em: <https://www.degruyter.com/view/books/9783110207576/9783110207576.1.27/9783110207576.1.27.xml>. Acesso em: 20 mar. 2019. Citado na p. 16.

MADDIESON, Ian. Tone. In: DRYER, Matthew S.; HASPELMATH, Martin (Ed.). **The World Atlas of Language Structures Online**. Leipzig: Max Planck Institute for Evolutionary Anthropology, 2013. Citado nas pp. 16, 35, 36.

MCGURK, Harry; MACDONALD, John. Hearing lips and seeing voices. **Nature**, v. 264, n. 12, p. 746–748, 1976. Citado na p. 12.

MCNEILL, David. Action, thought and language. en. **Cognition**, v. 10, n. 1-3, p. 201–208, jul. 1981. ISSN 00100277. DOI: 10.1016/0010-0277(81)90047-0. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/0010027781900470>. Acesso em: 29 ago. 2019. Citado nas pp. 12, 13.

MERMELSTEIN, P. Determination of the Vocal-Tract Shape from Measured Formant Frequencies. en. **The Journal of the Acoustical Society of America**, v. 41, n. 5, p. 1283–1294, mai. 1967. ISSN 0001-4966. DOI: 10.1121/1.1910470. Disponível em: <http://asa.scitation.org/doi/10.1121/1.1910470>. Acesso em: 11 set. 2019. Citado na p. 13.

MIXDORFF, Hansjorg; HU, Yu; BURNHAM, Denis. Visual Cues in Mandarin Tone Perception. en, p. 4, 2005. Citado nas pp. 16–19, 89, 100, 102.

NORTHERN DIGITAL MEASUREMENT SCIENCES. **Optotrak Certus**. Inglês. 2020. Disponível em: <https://www.ndigital.com/msci/products/optotrak-certus/>. Acesso em: 27 jun. 2020. Citado nas pp. 14, 46.

_____. **Optotrak Data Acquisition Unit (ODAU)**. Inglês. 2020. Disponível em: <https://www.ndigital.com/msci/products/optical-accessories/>. Acesso em: 27 jun. 2020. Citado na p. 46.

_____. **Vox-EMA**. Inglês. 2020. Disponível em: <https://www.ndigital.com/msci/products/vox-ema/>. Acesso em: 27 jun. 2020. Citado na p. 14.

OGLE, Derek H.; WHEELER, Powell; DINNO, Alexis. **FSA: Fisheries Stock Analysis**. 2020. R package version 0.8.30. Disponível em: <https://github.com/droglenc/FSA>. Citado na p. 84.

OHALA, John J. Production of Tone. In: FROMKIN, V. A. (Ed.). **Tone: A linguistic survey**. New York: Academic Press, 1978. Citado nas pp. 26–29, 37, 38, 44.

_____. The Relation between Phonetics and Phonology. In: THE Handbook of Phonetic Sciences. John Wiley & Sons, Ltd, 2010. P. 653–677. ISBN 978-1-4443-1725-1. DOI: 10.1002/9781444317251.ch17. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781444317251.ch17>. Acesso em: 7 fev. 2020. Citado na p. 39.

PAPOULIS, Athanasios. **Probability, Random Variables, and Stochastic Processes**. 4. ed.: McGraw-Hill, 2002. Citado na p. 54.

PERKELL, J. S. et al. Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. eng. **The Journal of the Acoustical Society of America**, v. 92, n. 6, p. 3078–3096, dez. 1992. ISSN 0001-4966. DOI: 10.1121/1.404204. Citado na p. 14.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria, 2019. Disponível em: <https://www.R-project.org/>. Citado nas pp. 83, 84.

RABINER, L.; SCHAFER, R. **Theory and Applications of Digital Speech Processing**. Pearson Education, 2011. ISBN 978-0-13-300253-9. Disponível em: <https://books.google.com.br/books?id=eS8rAAAAQBAJ>. Citado nas pp. 31, 55, 56.

SALTZMAN, Elliot L.; MUNHALL, Kevin G. A Dynamical Approach to Gestural Patterning in Speech Production. en. **Ecological Psychology**, v. 1, n. 4, p. 333–382, dez. 1989. ISSN 1040-7413, 1532-6969. DOI: 10.1207/s15326969eco0104_2. Disponível em: <http://www.tandfonline.com/doi/abs/10.1207/s15326969eco0104_2>. Acesso em: 4 set. 2019. Citado na p. 12.

SHAW, Jason A. et al. Influences of Tone on Vowel Articulation in Mandarin Chinese. en. **Journal of Speech, Language, and Hearing Research**, v. 59, n. 6, dez. 2016. ISSN 1092-4388, 1558-9102. DOI: 10.1044/2015_JSLHR-S-15-0031. Disponível em: <http://pubs.asha.org/doi/10.1044/2015_JSLHR-S-15-0031>. Acesso em: 11 fev. 2020. Citado na p. 45.

SHAW, Jason A et al. Vowel Identity Conditions the Time Course of Tone Recognition. en, p. 5, 2013. Citado na p. 45.

SILVA, Thaïs Cristófaro. **Fonética e fonologia do português: roteiro de estudos e guia de exercícios**. 7. ed. São Paulo: Contexto, 2003. ISBN 85-7244-102-6. Citado na p. 21.

SILVA, Thaïs Cristófaro et al. **Fonética Acústica: os sons do português brasileiro**. São Paulo: Contexto, 2019. ISBN 978-85-520-0079-2. Citado nas pp. 20, 21, 24, 25, 31–34.

SINGH, Leher; FU, Charlene S. L. A New View of Language Development: The Acquisition of Lexical Tone. en. **Child Development**, v. 87, n. 3, p. 834–854, mai. 2016. ISSN 00093920. DOI: 10.1111/cdev.12512. Disponível em: <http://doi.wiley.com/10.1111/cdev.12512>. Acesso em: 28 jan. 2020. Citado na p. 16.

SLUIJTER, Agaath M. C.; HEUVEN, Vincent J. van. Spectral balance as an acoustic correlate of linguistic stress. en. **The Journal of the Acoustical Society of America**, v. 100, n. 4, p. 2471–2485, out. 1996. ISSN 0001-4966. DOI: 10.1121/1.417955. Disponível em: <http://asa.scitation.org/doi/10.1121/1.417955>. Acesso em: 5 fev. 2020. Citado na p. 36.

SMITH, Damien; BURNHAM, Denis. Facilitation of Mandarin tone perception by visual speech in clear and degraded audio: Implications for cochlear implants. **The Journal of the Acoustical Society of America**, v. 131, n. 2, p. 1480–1489, fev. 2012. Publisher: Acoustical Society of America. ISSN 0001-4966. DOI: 10.1121/1.3672703. Disponível em: <https://asa-scitation-org.ez27.periodicos.capes.gov.br/doi/full/10.1121/1.3672703>. Acesso em: 1 abr. 2020. Citado nas pp. 16–18, 102.

SOARES, Marília Facó. Alguns Processos Fonológicos em Tükuna. pt. **Cadernos de Estudos Lingüísticos**, v. 10, p. 97–138, 1986. ISSN 2447-0686. DOI: 10.20396/cel.v10i0.8636720. Disponível em: <https:

//periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8636720>. Acesso em: 20 mar. 2019. Citado na p. 35.

SUMBY, W. H.; POLLACK, Irwin. Visual Contribution to Speech Intelligibility in Noise. en. **The Journal of the Acoustical Society of America**, v. 26, n. 2, p. 212–215, mar. 1954. ISSN 0001-4966. DOI: 10.1121/1.1907309. Disponível em: <http://asa.scitation.org/doi/10.1121/1.1907309>. Acesso em: 29 ago. 2019. Citado na p. 12.

TIEDE, Mark et al. Speech articulator movements recorded from facing talkers using two electromagnetic articulometer systems simultaneously. en. v. 11, p. 10, 2012. Citado na p. 14.

VATIKIOTIS-BATESON, Eric; MUNHALL, Kevin G. et al. Physiology-based synthesis of audiovisual speech. In: citado na p. 15.

VATIKIOTIS-BATESON, Eric; OSTRY, David J. An analysis of the dimensionality of jaw motion in speech. en. **Journal of Phonetics**, v. 23, n. 1, p. 101–117, jan. 1995. ISSN 0095-4470. DOI: 10.1016/S0095-4470(95)80035-2. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095447095800352>. Acesso em: 28 jan. 2020. Citado na p. 14.

VILELA BARBOSA, Adriano et al. Quantifying time-varying coordination of multimodal speech signals using correlation map analysis. en. **The Journal of the Acoustical Society of America**, v. 131, n. 3, p. 2162–2172, mar. 2012. ISSN 0001-4966. DOI: 10.1121/1.3682040. Disponível em: <http://asa.scitation.org/doi/10.1121/1.3682040>. Acesso em: 27 jan. 2020. Citado na p. 15.

WARD, I.C. **An Introduction to the Yoruba Language**. W. Heffer, 1952. Disponível em: <https://books.google.com.br/books?id=H39kAAAAMAAJ>. Citado na p. 35.

WEINBERG, Zach. **Svm separating hyperplanes**. 2012. Disponível em: <https://upload.wikimedia.org/wikipedia/commons/b/b5/Svm_separating_hyperplanes_%28SVG%29.svg>. Citado na p. 78.

WHALEN, D. H.; LEVITT, Andrea G. The universality of intrinsic F0 of vowels. en. **Journal of Phonetics**, v. 23, n. 3, p. 349–366, jan. 1995. ISSN 0095-4470. DOI: 10.1016/S0095-4470(95)80165-0. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095447095801650>. Acesso em: 11 fev. 2020. Citado na p. 44.

YEHIA, Hani C.; KURATATE, Takaaki; VATIKIOTIS-BATESON, Eric. Linking facial animation, head motion and speech acoustics. en. **Journal of Phonetics**, v. 30, n. 3, p. 555–568, jul. 2002. ISSN 0095-4470. DOI: 10.1006/jpho.2002.0165. Disponível em:

<http://www.sciencedirect.com/science/article/pii/S0095447002901658>. Acesso em: 11 fev. 2020. Citado nas pp. 45, 98.

YEHIA, Hani; RUBIN, Philip; VATIKIOTIS-BATESON, Eric. Quantitative association of vocal-tract and facial behavior. en. **Speech Communication**, v. 26, n. 1-2, p. 23–43, out. 1998. ISSN 01676393. DOI: 10.1016/S0167-6393(98)00048-X. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S016763939800048X>. Acesso em: 11 set. 2019. Citado nas pp. 13, 15.

YIP, M. **Tone**. Cambridge University Press, 2002. (Cambridge Textbooks in Linguistics). ISBN 978-0-521-77445-1. Disponível em: <https://books.google.com.br/books?id=KFv2lojXjpwC>. Citado nas pp. 16, 17, 26, 29, 35–44, 54, 100.

YIP, Moira Jean. **The tonal phonology of Chinese**. 1980. Thesis – Massachusetts Institute of Technology. Disponível em: <http://dspace.mit.edu/handle/1721.1/15971>. Acesso em: 4 abr. 2019. Citado na p. 35.

YUAN, Jiahong. Perception of intonation in Mandarin Chinese. **The Journal of the Acoustical Society of America**, v. 130, n. 6, p. 4063–4069, dez. 2011. ISSN 0001-4966. DOI: 10.1121/1.3651818. Disponível em: <https://asa-scitation-org.ez27.periodicos.capes.gov.br/doi/abs/10.1121/1.3651818>. Acesso em: 20 mar. 2019. Citado na p. 16.

ZAHNER, Marlene et al. Conversion from Facial Myoelectric Signals to Speech: A Unit Selection Approach. en, p. 5, 2014. Citado na p. 14.

ZEE, Eric. Tone and vowel quality. en. **Journal of Phonetics**, v. 8, n. 3, p. 247–258, jul. 1980. ISSN 0095-4470. DOI: 10.1016/S0095-4470(19)31474-3. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0095447019314743>. Acesso em: 11 fev. 2020. Citado na p. 44.