UNIVERSIDADE FEDERAL DE ALAGOAS
CAMPUS A. C. SIMÕES
FACULDADE DE LETRAS
PROGRAMA DE PÓS-GRADUAÇÃO EM LINGUÍSTICA E LITERATURA

JULIO CESAR GALDINO

EM 200 METROS, VIRE À ESQUERDA: A ENTOAÇÃO DOS COMANDOS DE GPS

Maceió

2023

JULIO CESAR GALDINO

EM 200 METROS, VIRE À ESQUERDA: A ENTOAÇÃO DOS COMANDOS DE GPS

Dissertação apresentada ao Programa de Pós-Graduação em Linguística e Literatura – PPGLL, da Universidade Federal de Alagoas – UFAL, como requisito parcial à obtenção do título de mestre em Linguística.

Linha de Pesquisa: Teoria e Análise Linguística.

Orientador: Prof. Dr. Miguel José Alves de Oliveira Júnior

Maceió

2023

Catalogação na fonte Universidade Federal de AlagoasBiblioteca Central Divisão de Tratamento Técnico

Bibliotecária: Taciana Sousa dos Santos – CRB-4 – 2062

G149d Galdino, Julio Cesar.

Em 200 metros, vire à esquerda : a entoação dos comandos de GPS/ Julio Cesar Galdino. – 2023.

86 f. : il.

Orientador: Miguel José Alves de Oliveira Júnior.

Dissertação (Mestrado em Linguística) – Universidade Federal de Alagoas. Programa de Pós-Graduação em Linguística e Literatura. Maceió, 2023.

Bibliografia: f. 66-80.

Apêndices: f. 81-86.

1. Sistema de Posicionamento Global (GPS). 2. Teoria dos atos de fala. 3. Prosódia. 4. Entoação. I. Título.

CDU: 81

Dedico esta dissertação a meus pais,
Hilda de Melo Galdino e Francisco Galdino Neto.

AGRADECIMENTOS

O trajeto desta dissertação ocorreu graças às instruções de muitos falantes.

Agradeço a Miguel Oliveira Jr., pelo convite ao grupo de pesquisa, pela sugestão do tema, pela orientação e por todo o apoio. Obrigado por acreditar no serviço público e por depositar sua confiança em pessoas iniciantes na pesquisa. Sem suas direções, esta pesquisa não teria chegado ao destino final dela, e eu não teria chegado tão longe e aprendido tanto.

Agradeço a meus pais, Hilda e Francisco, por sempre terem entendido que esta trajetória era importante para mim e por terem me mostrado caminhos que vão além do percurso acadêmico.

Agradeço a todos da banca de defesa e de qualificação por terem aceitado o convite e pelas sugestões que fizeram a versão final deste trabalho ser mais precisa e confiável.

Agradeço as duas participantes pelas vozes cedidas para produzir as rotas neste idioma português brasileiro na variedade alagoana.

Agradeço a todos do grupo FonUFAL, que, apesar da distância, mostrou-se muito acessível, além de interativo e eficiente.

Esta pesquisa teria um fluxo muito mais difícil se não fosse a agradabilidade de Maryanne, as coordenadas que Arthur Terto me forneceu do Praat, o suporte que Eliel e Alice me deram nas anotações dos dados, a solicitude de Kyvia Fernanda no início do mestrado, a disponibilidade de Ayane e René e os cursos acadêmicos mapeados pelo Thiago Vidotto. Agradeço também a Miguel, Maryanne, Arthur, Cris, Ebson, Marvin, Selma e Ana, por unirem inteligência e humor, essenciais para um tráfego acadêmico tranquilo.

Por todas as vezes que me disseram para seguir em frente, agradeço a Bruna Oliveira, Israel Faustino, Anderson Ferreira, Felipe Jackson e demais amigos que foram meus pontos de referência para que eu não me perdesse.

RESUMO

O GPS (*Global Positioning System*) ou Sistema de Posicionamento Global é uma tecnologia que utiliza uma voz produzida por síntese de fala para fornecer os comandos de orientação e instruir os ouvintes. Por se tratar de uma instrução, a fala do GPS se enquadra na Teoria dos Atos de Fala, mais especificamente nos atos diretivos, que têm como objetivo fazer com que o ouvinte realize a ação proferida pelo falante. Estudos apontam o importante papel da prosódia na fala diretiva, uma vez que a entoação tem a função de diferenciar, na comunicação, os atos de fala diretivos (SANTOS, 2010; GOMES DA SILVA, CARNAVAL & MORAES, 2020). Embora os sistemas de síntese não apresentem problemas no nível segmental, estudos têm demonstrado que ainda falham no nível prosódico (TAYLOR, 2009). O presente estudo propôs analisar e descrever os padrões entoacionais de falas diretivas produzidas por sistemas de GPS e compará-los com a voz humana. Para atingir esse objetivo, o corpus de análise utilizou 20 frases do *Google Maps*, 20 frases do *Waze* e a realização dessas frases por duas participantes do sexo feminino, nativas do português brasileiro na variedade alagoana. A análise acústica e a descrição das curvas foram feitas com os *scripts* AnalyseTier (HIRST, 2012) e ProsodyPro (XU, 2013), em conjunto com o Praat, e aplicamos testes estatísticos por meio do software R. Os resultados mostraram que houve diferenças na entoação de aplicativos e humanos ao avaliar a variação de f0, além de diferenças acústicas significativas nas médias de f0 média e f0 máxima em todas as unidades entoacionais analisadas. Os resultados também mostraram semelhanças entre aplicativos e humanos, na configuração de f0, embora haja diferenças ao verificar a representação gráfica das curvas. Tais resultados reforçam o avanço da síntese de fala, dadas as semelhanças na configuração de f0, no entanto, também indicam que ainda é necessário aprimorar esses sistemas, uma vez que houve diferenças significativas entre essa fala e a voz natural.

Palavras-chaves: GPS; entoação; fala diretiva.

ABSTRACT

GPS (Global Positioning System) is a technology that uses a voice produced by speech synthesis to provide orientation commands and instruct listeners. As it is an instruction, GPS speech falls within Speech Act Theory, more specifically in directive acts, which aim to make the listener perform the action uttered by the speaker. Studies point to the important role of prosody in directive speech, since intonation has the function of differentiating directive speech acts in communication (SANTOS, 2010; GOMES DA SILVA, CARNAVAL & MORAES, 2020). Although synthesis systems do not present problems at the segmental level, studies have shown that they still fail at the prosodic level (TAYLOR, 2009). The present study proposed to analyze and describe the intonational patterns of directive speeches produced by GPS systems and compare them with the human voice. To achieve this goal, the analysis corpus used 20 phrases from Google Maps, 20 phrases from Waze and the realization of these phrases by two female participants, native speakers of Brazilian Portuguese in the Alagoan variety. Acoustic analysis and description of curves were done with AnalyseTier scripts (HIRST, 2012) and ProsodyPro (XU, 2013), in conjunction with Praat, and we applied statistical tests using R software. The results showed that there were differences in intonation between applications and humans when evaluating f0 variation, as well as significant acoustic differences in mean f0 and maximum f0 averages in all intonational units analyzed. The results also showed similarities between applications and humans in f0 configuration, although there are differences when checking the graphical representation of curves. Such results reinforce the advancement of speech synthesis given the similarities in f0 configuration; however, they also indicate that it is still necessary to improve these systems since there were significant differences between this speech and natural voice.

Keywords: GPS; intonation; directive speech.

LISTA DE FIGURAS

LISTA DE GRÁFICOS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Autossegmental métrico
ANOVA	Analysis of Variance
ASR	Automatic Speech Recognition
CEP	Comitê de Ética e Pesquisa
COS	Center for Open Science
CWT	Continuous Wavelet Transform
DaTo	Dynamic Tones of Brazilian
DCT	Discourse Completion Test
DNN	Deep Neural Networks
HMM	Hidden Markov Model
Hz	Hertz
F0	Frequência fundamental
GM	Google Maps
GPS	Global Position System
HUM	Humanos
OSF	Open Science Framework
P1	Participante 1
P2	Participante 2
PB	Português Brasileiro
pUE	Primeiro pico da unidade entoacional
RNN	Rede Neural Recorrente
TAF	Teoria dos Atos de Fala
TCLE	Termo de Consentimento Livre e Esclarecido
TTS	Text-to-speech

ToBI	Tone and Break Indices
UE	Unidade Entoacional
UFAL	Universidade Federal de Alagoas
WZ	Waze

SUMÁRIO

1 INTRODUÇÃO

O GPS (*Global Positioning System*) ou Sistema de Posicionamento Global é um mecanismo que permite encontrar posições geográficas e utiliza uma fala sintética para fornecer os comandos de orientação. Essa voz tem como característica instruir os ouvintes, a fim de que eles sigam rotas e cheguem ao seu destino. Por se tratar de uma instrução, a fala do GPS se enquadra na Teoria dos Atos de Fala, mais especificamente nos atos diretivos, que têm como objetivo fazer com que o ouvinte realize a ação proferida pelo falante.

Os navegadores de GPS permitem a transmissão de informações e o direcionamento geográfico com o auxílio de uma voz produzida por síntese de fala. Essa produção pode ser realizada por síntese de fala concatenada, que é a reprodução de amostras pré-gravadas, com sentenças ou palavras simples da voz natural, ou pode ser o resultado de síntese de formantes, ou seja, uma modelagem da fonte e das frequências do som (FILHO, 2002).

As frases proferidas pelo GPS normalmente funcionam como um guia, em sentenças como "Vire à direita". Dessa maneira, esses tipos de enunciados estão inseridos no âmbito da Teoria dos Atos de Fala (AUSTIN, 1990; SEARLE, 1991), nos atos ilocutórios de fala diretiva (pedir, ordenar, aconselhar etc.).

Para garantir um grau razoável de inteligibilidade e aceitabilidade, é importante que a voz dos sistemas de síntese de fala soe o mais natural possível (SOUZA, 2010). Em geral, embora esses sistemas não apresentem problemas no que diz respeito ao nível segmental, como produções adequadas de palavras, estudos têm demonstrado que ainda falham no nível prosódico (TAYLOR, 2009). Se a fala sintetizada do GPS for realizada de uma maneira inadequada, isso pode gerar prejuízos para a comunicação.

Alguns estudos apontam o importante papel da prosódia na fala diretiva. A entoação, por exemplo, tem a função de diferenciar, na comunicação, os atos de fala diretivos (SANTOS, 2010; GOMES DA SILVA, CARNAVAL & MORAES, 2020), uma vez que diferentes elementos dos correlatos físicos demonstram diferentes atitudes do locutor (QUEIROZ, 2011). Além disso, os atos de fala diretivos costumam ser reduzidos ao modo imperativo, mas eles não apresentam diferenças na estrutura sintática em relação à intenção comunicativa, tendo a prosódia um papel modalizador, capaz de diferenciar ordem, súplica, convite, instrução, entre outros (BODOLAY, 2009).

Para Rocha (2013), a instrução possui uma determinada entoação para o português brasileiro. Então, analisar a voz do GPS e compará-la com a voz humana é importante, pois é

necessário verificar semelhanças e diferenças entre elas. Por vezes, esses sistemas demonstram ter uma entoação distante da fala humana, o que se torna um problema, pois pode haver ruídos na comunicação que afetem a percepção do usuário, por exemplo. A presente dissertação parte da hipótese de que os padrões entoacionais realizados pela voz do GPS são diferentes da voz humana em relação à frequência fundamental.

O GPS é uma ferramenta bastante utilizada no cotidiano, cuja compreensão dos conteúdos enunciados tem impacto direto nas decisões tomadas no trânsito e nas ruas, por exemplo. Portanto, validar se a fala produzida por esses sistemas se aproxima daquilo que se tem descrito como característico da fala natural é um empreendimento de importância, não apenas científica, mas também social.

O presente estudo tem como objetivo geral descrever os padrões entoacionais de falas diretivas produzidas por sistemas de GPS. Entre os objetivos específicos, estão: (i) analisar acusticamente contornos de frequência fundamental de falas diretivas produzidas por dois sistemas populares de GPS (*Google Maps* e *Waze*), (ii) descrever, a partir dessas análises, os padrões entoacionais desses sistemas, (iii) analisar acusticamente contornos de frequência fundamental dessas falas diretivas do GPS produzidas por duas falantes do português na variedade alagoana, (iv) comparar os padrões entoacionais dos sistemas de GPS com os padrões das duas falantes, a partir de frases idênticas, e (v) comparar esses resultados com padrões entoacionais típicos de falas diretivas tal como descritos pela literatura.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo disserta sobre alguns aspectos da prosódia, especificamente sobre a entoação, e apresenta também uma revisão a respeito da síntese de fala e como as informações prosódicas são tratadas nas vozes dos sistemas que utilizam essa voz sintética.

Além disso, foi preciso compreender a Teoria dos Atos de Fala, os atos diretivos e também como a prosódia possui papel importante para diferenciar esse tipo de ato, sobretudo a instrução.

2.1 Prosódia e entoação

Prosódia e entoação são termos que, com frequência, são utilizados como equivalentes na literatura (MUSILIYU; OLIVEIRA JR., 2015). Para Hirst e Di Cristo (1998), as diferenças de uso dos dois termos devem-se a uma dupla ambiguidade quando se fala em entoação. A primeira ambiguidade, conforme os autores, é o fato de considerar ou não a entoação como um fato mais restrito em relação à prosódia, enquanto que a segunda estaria relacionada a uma distinção na forma de descrever os fatos.

Couper-Kuhlen (1986) apresenta um modelo cuja entoação é tratada no sentido restrito da melodia da fala, envolvendo características prosódicas ligadas sobretudo ao tom (direção) e ao alcance do tom (altura e largura), e a prosódia envolveria traços, como sonoridade, ritmicidade, pausa etc. Se considerarmos a literatura fonológica sobre a prosódia, a entoação é um dos aspectos prosódicos das línguas, assim como a duração e o acento (MATEUS, 2005; MATEUS *et al.*, 2003). A entoação, neste estudo, será utilizada em um sentido mais restrito, a partir da análise da frequência fundamental (F0).

A F0 corresponde ao número de vezes em que há vibração das pregas vocais e tem seu valor medido em Hertz (Hz) (BARBOSA, 2019). Os valores de F0 podem também ser expressos em semitons (relativos a 100Hz), que geralmente são utilizados para descrever medidas de variabilidade, uma vez que o Hertz não representa a percepção do tom com precisão e apresenta uma visão distorcida da variabilidade de uma voz (CUMBERS, 2013). O correlato perceptual da f0 é o *pitch*, medido em semitons, mas é importante notar que essa medida não cresce de forma logarítmica com essa frequência, fazendo com que a sensação de percepção seja distinta da f0 (BARBOSA, 2019). A f0 emite taxas mínimas, médias, máximas e outras variantes que permitem caracterizar a vibração das pregas vocais (CAVALCANTE, 2009).

A entoação é descrita a partir da análise da curva de F0 (ou contorno melódico), apesar de variações de duração também participarem na expressão de termos atribuídos à entoação (GOMES DA SILVA *et. al*, 2020). Moraes & Rilliard (2022) explicam que as curvas de F0 são obtidas por programas de acústica da fala, como o Praat (BOERSMA; WEENINK, 1993-2023), a partir de processos de suavização dos contornos ou por um processo mais radical, de estilização, que mostra os contornos mediante segmentos de reta.

No fluxo da fala, há modulações melódicas que são caracterizadas por mudanças de direção da F0 (ascendentes ou descendentes) ou "tons de fronteira" (MORAES; RILLIARD, 2022). Além dessas direções, os contornos melódicos podem ser planos, ou seja, em que não há um movimento ascendente ou descendente significativo na sentença (SANTOS, 2016).

Conforme Moraes & Rilliard (2022), uma das mais notórias funções da entoação é a função ilocutória, que caracteriza os atos de fala, ou atos ilocutórios, sendo que o português possui um sistema entoacional bastante rico nesse sentido. Estudos envolvendo entoação e fala diretiva têm empregado o modelo autossegmental-métrico (AM) para descrever os padrões entoacionais dos atos de fala (BARRETO, 2019; GOMES DA SILVA *et. al*, 2020). O (AM) é uma teoria que tem por objetivo fonológico descrever contornos em uma série de caracteres categoricamente distintos e tem por objetivo fonético mapear elementos fonológicos para parâmetros acústicos contínuos, resolvendo questões gerais de estrutura suprassegmental e descrevendo contornos de entoação (LADD, 2008).

Essa teoria usa símbolos representados por letras que indicam um tom alto (H – *high*), um tom baixo (L – *low*), ou uma realização bitonal, como L*+H (FROTA & MORAES, 2016; FROTA *et al*., 2015). Esses rótulos podem ser acompanhados por diacríticos que indicam (-) acento frasal, (%) tom de fronteira, (*) acento tonal, (!) início da compressão da gama tonal e (?) incerteza de algum tom ou de um tipo de tom (SVARTMAN, 2021). Trabalhos que descreveram o Português Brasileiro (PB) mostram que os acentos tonais podem ser (L* ou H*), ou seja, monotonais, ou podem ser bitonais (FROTA & VIGÁRIO, 2000; FROTA *et al*., 2015; SERRA, 2009; TENANI, 2002).

Esse tipo de representação entoacional deu origem ao ToBI (*Tone and Break Indices*), desenvolvido para o inglês americano padrão (BECKMAN; HIRSCHBERG, 1994), baseado em trabalhos de outros pesquisadores (BECKMAN & PIERREHUMBERT, 1986; LADD, 1993; PIERREHUMBERT, 1980; PRICE *et al*., 2015; WIGHTMAN *et al*., 1992). Para o português, há o P_ToBI (FROTA *et al*., 2015) e o DaTo (*Dynamic Tones of Brazilian Portuguese*) proposto por Lucente (2012). Essas descrições entoacionais podem ser feitas

manualmente, porém, existem *scripts* utilizados em programas de análise acústica da voz que são capazes de fazer uma anotação entoacional de forma automática, como o Momel/Intsint (HIRST, 2007).

2.2 A conversão de texto para a fala – TTS (*Text-to-Speech*)

A síntese de fala é a produção de voz por máquinas, a partir da fonetização automática de frases (DUTOIT, 1997). Ao contrário da simples reprodução de voz, essa síntese objetiva o equivalente à produção da fala humana, com informações fonéticas e prosódicas correspondentes (SAGISAKA, 1990).

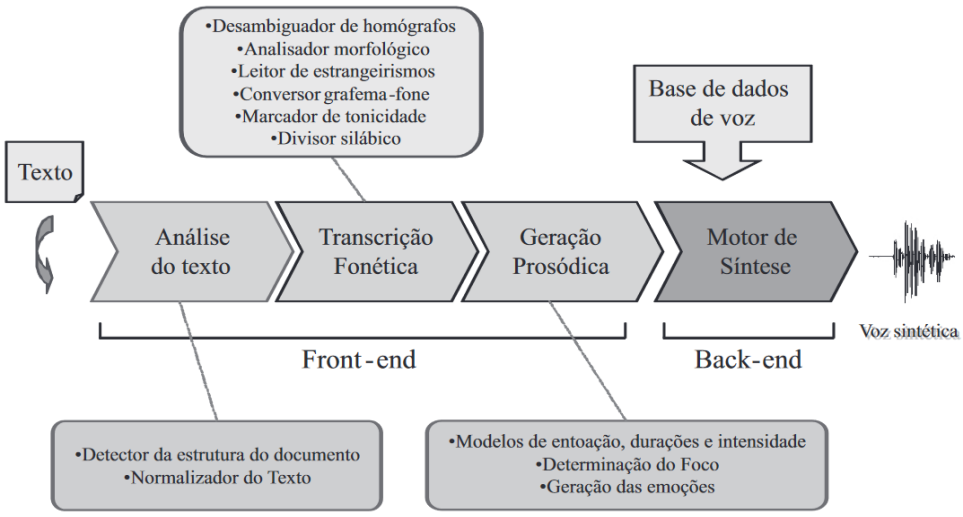
A síntese de fala se divide em duas categorias e são diferenciadas a partir do tamanho do vocabulário e do campo de aplicação, conforme Egashira (1992). Na primeira categoria, estão os sistemas de resposta vocal, usados em serviços telefônicos, sistemas de saldo bancário, por exemplo, com frases introdutórias, como "bom dia", "digite sua senha", em que há pouca interação com o usuário. Nesses casos, o vocabulário é limitado, e sua realização é resultado de gravação e armazenamento de fala, a fim de gerar possibilidades de combinações para uma posterior reprodução.

A segunda categoria dos sistemas de síntese de fala são os chamados conversores de texto em fala (TTS). Eles contêm uma gama enorme de aplicações, facilitando a interação humano-computador para cegos, lendo notícias, boletins meteorológicos e, principalmente, atuando na automação de *call center* (TAYLOR, 2009). Assim, essa classe de sistema possui um vocabulário irrestrito, tem um custo computacional mais elevado e precisa fazer análises do texto escrito, identificação dos sons equivalentes, associações dos parâmetros de entoação e ritmo e processamento de sinais, o que não traz, muitas vezes, a naturalidade que a fala humana possui (PACHECO, 2010).

Os modernos sistemas de TTS possuem diversas arquiteturas, mas existem pelo menos três blocos que são comuns: *front-end* (pré-processamento de texto), *back-end* (motor de síntese) e *voice font* (base de dados de voz) (BRAGA, 2007). O *front-end* contém as seguintes etapas: (1) análise de texto, que separa os limites de cada frase, normalizando o texto, e transformando símbolos, abreviações, siglas em segmentos pronunciáveis, como proferir "cinquenta reais" (50 reais) e não "cinco zero reais" (50 reais); (2) análise fonética, que converte os grafemas em fones e que desambigua homônimos, como em "Eu gosto de pão", em que o sistema deve perceber que é "gósto" e não "gôsto", por exemplo; e (3) geração de prosódia, em que se estima *pitch*, intensidade e duração (SÁ, 2018). Depois que o *front-end* interpreta a

transcrição fonética, o motor de síntese contido no *back-end* a transforma em fala sintética (BRAGA, 2007). Um resumo dessa arquitetura tradicional é mostrado na figura 1.

Figura 1 - Arquitetura tradicional de um sistema de conversão texto-fala



Fonte: Braga (2007)

As etapas de análise do texto e de transcrição fonética do *front-end* não apresentam grandes dificuldades. Entretanto, a última etapa, a geração prosódica, ainda enfrenta desafios, pois os sistemas de síntese carecem muitas vezes de naturalidade, resultando nesses casos em uma artificialidade da voz sintética (SILVA, 2004). De acordo com Klimkov *et al*. (2017), o investimento na anotação prosódica pode resultar no melhoramento da naturalidade da fala sintética.

Além dos sistemas TTS, outro tipo de aplicação tecnológica que se ampara em informações prosódicas da fala são os sistemas ASR (*Automatic Speech Recognition*), em que, ao contrário da síntese de fala, a voz é inserida no sistema e é convertida em texto. Esses sistemas também usam um *front-end*, mas para converter a voz em uma matriz X de parâmetros, em um modelo probabilístico, ou em uma linguagem não probabilística, em modelos de gramática livre de contexto (NETO *et al.*, 2005). Segundo Tevah (2006), o uso de ASR pode ser distribuído em sistemas de interface e em sistemas transcritores: o primeiro grupo permite acelerar ações de comando, como os celulares que discam por meio dos comandos de voz, enquanto os sistemas transcritores possuem requisitos técnicos maiores, transcrevendo o que o usuário estiver falando, como se fosse uma secretária virtual. A informação prosódica pode ajudar a melhorar esses sistemas, uma vez que recursos, como a pausa, a intensidade, o *pitch* e a frequência fundamental, podem possibilitar o reconhecimento de voz por meio de redes neurais (BALLESTEROS & WANNER, 2016; LIU, LIU & SONG, 2018; SZASZÁK &

TÜNDIK, 2019; YI & TAO, 2019). A inteligência artificial já apresenta bons resultados no reconhecimento dessa fala, mas há escassez de trabalhos em língua portuguesa (TEIXEIRA *et al*, 2018).

2.2.1 TTS e prosódia

Muitos sistemas TTS predizem representações prosódicas diretamente do texto, mas há risco de o processo de análise cometer erros, suscitando o desafio de gerar conteúdo prosódico, uma vez que não há como descobrir todas as informações prosódicas somente a partir do texto (TAYLOR, 2009). Essa preocupação com a prosódia é uma constante em estudos sobre TTS. Trabalhos sobre o português brasileiro, por exemplo, descrevem os procedimentos desses sistemas, abordando, também, diferentes conceitos relativos à produção da fala, como fones, duração, ritmo, frequência fundamental, intensidade etc. (EGASHIRA, 1992; GOMES, 1998; SILVA; VIOLARO, 1995; OLIVEIRA, 1996; CHBANE, 1994; SIMÕES, 1999; BARBOSA, 1999).

Um dos parâmetros prosódicos mais investigados nesses estudos é a frequência fundamental (F0). Nesse sentido, os trabalhos sobre síntese de fala começaram a propor modelos para gerar frequência fundamental automaticamente, a partir de informações obtidas em bancos de dados de fala natural, para tentar se aproximar dos contornos de F0 da fala humana. Esses modelos usam em geral a técnica de concatenação de contorno para gerar a frequência fundamental (DUSTERHOFF & BLACK, 1997; SILVA, 1998; SAITO, 1998).

A utilização de redes neurais artificiais, a fim de melhorar a voz sintetizada, é utilizada desde o século passado. Essas redes conseguem aprender automaticamente regras fonológicas da fala e são aplicadas com sucesso na geração de prosódia (SONNTAG, 1997; CHEN *et al.*, 1999; SCORDILIS & GOWDY, 1989).

No século XXI, percebe-se a divisão das técnicas de síntese da fala em três gerações: síntese baseada em regras, síntese concatenativa e síntese por modelo baseado em HMMs - *Hidden Markov Models* (TAYLOR, 2009). Segundo Leite *et al.* (2014), os métodos mais usados atualmente são a síntese concatenativa e a síntese HMMs, sendo que há vantagens e desvantagens em ambas. A síntese baseada em Modelos Ocultos de Markov facilita a alteração da voz artificial, a partir da elaboração da prosódia do discurso e das características da voz, mas deixa a fala menos natural, porque concatena modelos estatísticos dos sons. Enquanto isso, a

síntese concatenativa baseia-se em segmentos de fala natural pré-gravada, o que gera, consequentemente, uma voz mais próxima da humana, mas usa um recurso limitado de dados.

Esses métodos são investigados, em sua maioria, por profissionais da Engenharia da Computação. No entanto, Simões *et al.* (2000) propõem um sistema TTS para o português brasileiro com colaboração de linguistas, o que aponta a necessidade de formação interdisciplinar de linguistas neste campo de investigação (BRAGA, 2007). Trabalhos na área da Linguística aplicados para o desenvolvimento de tecnologias da fala apresentam, em geral, abordagem da fonética acústica, considerando as variantes dialetais para o desenvolvimento de TTS ou analisando a fala em aplicativos (PAIVA *et al.*, 2004; CORREA, 2010; MANFIO, 2012). Em um desses estudos, Manfio (2012) comenta a síntese de voz à luz de teorias linguísticas, acerca de um aplicativo disponível chamado *Balabolka*. O autor explica que o app se utiliza de partes gravadas da voz humana, cria as palavras que pronuncia, possui uma prosódia próxima à da fala real nas frases afirmativas, mas não consegue ainda atingir um bom desempenho nas frases interrogativas e na realização dos ditongos.

Nos últimos anos, os estudos em português mantiveram a preocupação de comparar métodos de síntese de TTS, propor blocos de processamento de voz ou analisar novas bases de vozes (BARBOSA, 2016; CASANOVA, 2019; MOREIRA, 2015; VECCHIETTI, 2015). Essas análises costumam se amparar em informações prosódicas, como em Maia & Seara (2017), que propõem um método de síntese de fala a partir do *pitch*, baseado em redes neurais profundas (*Deep Neural Networks*, DNN).

Estudos em outras línguas, como o inglês, o chinês e o tailandês, por exemplo, apresentam uma variedade maior de análises, com o objetivo principal de propor modelos para geração de contornos de frequência fundamental (KAMEOKA *et al.*, 2015; KORIYAMA & KOBAYASHI, 2015; THOMAS *et al.*, 2015; MOUNGSRI *et al.*, 2017; CHEN *et al.*, 2018; RAO, 2017). Em uma dessas análises, Ronanki (2019) propõe um método de geração automática de F0 no inglês britânico, tendo como resultado uma melhora na F0, a partir da geração de contornos de *pitch* previstos por uma rede neural recorrente (RNR). Em línguas como o mandarim, a F0 baseia-se em "tons" lexicais que diferem em significado, havendo necessidade de um bom resultado dos padrões entoacionais (CHEN *et al.*, 2018). Assim, definir esses contornos é importante, pois uma prosódia que se distancia das características da fala natural pode prejudicar a inteligibilidade dos sistemas.

A geração da frequência fundamental (F0) desempenha um papel importante na síntese da fala, uma vez que influencia diretamente na naturalidade da fala sintética (JANYOI;

SERESANGTAKUL, 2020b). O modelo de processo de geração de contornos de F0 é capaz de representar bem a estrutura hierárquica da prosódia, que contém intervalos de tempo curtos, cobrindo sílaba ou palavra, e intervalos de tempo mais longos, cobrindo frases e parágrafos (HIROSE, 2016). Novos métodos são sempre propostos e têm se mostrado promissores para a geração de contorno F0 de alta qualidade ao mesmo tempo em que fazem uso eficiente dos dados (LANGARANI; SANTEN, 2016).

Esses modelos de f0 são geralmente baseados em redes neurais, que permitem uma melhoria na qualidade da voz (WANG *et al.*, 2017; JANYOI & SERESANGTAKUL, 2020a; KLIMKOV *et al.*, 2018; KLIMKOV *et al.*, 2019; LATORRE *et al.*, 2019; YIN *et al.*, 2016). Essas redes são utilizadas juntamente com as técnicas de aprimoramento de F0 nos sistemas. Uma técnica recentemente proposta é a Transformada Wavelet Contínua (CWT - *Continuous Wavelet Transform*), que descreve o sinal de F0 e tenta, por exemplo, converter uma fala neutra em uma fala emocional (MING *et al.*, 2015; RIBEIRO *et al.*, 2015; ZHAO *et al.*, 2020). Esses modelos de redes neurais podem capturar duração implícita, f0 e intensidade, recursos prosódicos que possuem um papel importante na melhoria da qualidade do sistema de síntese de texto para fala (REDDY; RAO, 2015).

Um tipo de método de síntese de fala bastante utilizado é a modelagem de contornos de f0 baseada em HMMs (GHONE *et al.*, 2017; WANG *et al.*, 2015; LANGARANI *et al.*, 2015; DALL *et al.*, 2016). Zen *et al.* (2009) explicam que a síntese baseada em HMMs (Modelos Ocultos de Markov) tem como vantagem a unificação dos blocos *front-end* e *back-end*, gerando uma nova estrutura. Esse método pode ser combinado com as redes neurais, mas, em alguns casos, pode levantar questões sobre como esses dois recursos interagem e como combiná-los corretamente (COOPER, 2019).

Abordagens em HMMs ainda superam os métodos de aprendizado profundo na precisão de predição de f0, devido ao seu comportamento descontínuo (TÓTH; CSAPÓ, 2016). Pesquisas têm utilizado bancos de dados de fala modelados por HMMs levando em conta contextos fonéticos e prosódicos (ZEN *et al.*, 2007). Tendo em vista a utilização desses contextos, a Linguística se torna importante para a geração da fala sintética, pois é a ciência que permite, entre outras análises, a descrição de aspectos articulatórios e prosódicos da fala, a partir de teorias.

Nos últimos anos, há uma tentativa de implantar expressões emocionais para melhorar a expressividade de TTS, usando, também, redes neurais (INOUE *et al.*, 2017; ROBINSON *et al.*, 2019; TAHON *et al.*, 2018). Os métodos atuais de conversão de texto em fala produzem

vozes com sons realistas, mas não têm um bom desempenho da expressividade emocional que os falantes esperam (ROBINSON *et al.*, 2019). Na fala emocional, algumas palavras e frases são pronunciadas com destaque se comparada à fala neutra, segundo Yadav & Rao (2015). Para os autores, o uso de fragmentos de nível de sílabas teve melhor desempenho do que os fragmentos de nível de palavra ou de frase para gerar emoção na voz dos sistemas.

O processo de geração de expressividade em TTS envolve normalmente a conversão de uma voz neutra para uma voz emocional, dependendo, por exemplo, de cálculos de interação do contexto da frase (ROBINSON, *et al.*, 2019). Dessa forma, tendo em vista a necessidade de considerar diversos aspectos linguísticos, a Linguística se faz importante, pois uma teoria voltada para o uso da linguagem pode contribuir para uma fala sintética mais expressiva.

2.3 A Teoria dos Atos de Fala

A Teoria dos Atos de Fala (TAF), proposta inicialmente por John Langshaw Austin, é conhecida por estudar as afirmações que correspondem à realização de uma ação. Para Austin (1990), há dois tipos de enunciados: as sentenças constativas, que descrevem estados de coisas, e as sentenças performativas, que se realizam na execução de uma ação ao emitir um proferimento.

Austin (1990) propõe que os proferimentos performativos são divididos em três atos de fala: (i) o ato locucionário (ou locucional), que se realiza ao enunciar uma frase, (ii) o ato ilocucionário (ou ilocucional), que se realiza na linguagem, e (iii) o ato perlocucionário (ou perlocucional) que se dedica aos estudos das perlocuções, ou seja, as consequências ocasionadas no interlocutor a partir do ato ilocucionário.

A TAF é retomada a partir dos estudos de John Searle, que desenvolve uma série de aspectos dessa teoria, mas tendo como foco a análise do ato ilocucionário. Segundo Searle (1991), esses atos possuem uma espécie de força ilocucional e vêm acompanhados de verbos que podem afirmar, perguntar ou mesmo ordenar.

Escandell-Vidal (1993) explica que o significado de qualquer oração pode ser analisado por um indicador proposicional, que é o assunto da sentença, e por um indicador de força ilocucional, que mostra qual o sentido, ou seja, com que força ilocutiva essa sentença deve ser interpretada.

Searle (1991) propõe que as ilocuções possuem uma forma geral ilustrada pela expressão *F (p)*, em que "F" significa a força ilocucional e "p" significa uma determinada proposição. A partir dessa explicação, é possível demonstrar diferentes tipos de ilocuções,

simbolizando *Pr(p)* para promessas, *!(p)* para pedidos, */-(p)* para asserções, *?(p)* para perguntas do tipo sim/não e *W(p)* para advertências. Dessa forma, uma frase como "Você fez isso?" seria simbolizada como "*?*(Você fez isso)".

É apenas em outra obra de Searle que os problemas mais relevantes da TAF são explicados e que os atos ilocucionários são classificados alternativamente em cinco categorias gerais de se usar a linguagem. Searle (1995) classifica como primeira categoria de ato ilocucionário as declarações assertivas, que podem ser avaliadas em verdadeiras e falsas. Dessa forma, por exemplo, verbos como "concluir" e "deduzir" denotam afirmativas com a característica de que tem algo a ver com o interesse do falante.

A segunda categoria dos atos é chamada de diretiva e são caracterizados pela tentativa do falante para fazer o ouvinte realizar alguma ação. São tentativas que podem ser mais modestas, como o convite ou a instrução, ou podem ser mais autoritárias, como a ordem. Os verbos que aparecem relacionados à fala diretiva, segundo Searle (1995), são perguntar, ordenar, comandar, solicitar, pedir, implorar, orar, suplicar, convidar, permitir e aconselhar. Além desses, o autor também inclui os verbos ousar e desafiar classificados como comportamentais por Austin (1990) e muitos dos verbos classificados como exercitivos.

A terceira categoria pertence aos atos comissivos, que têm por objetivo comprometer o falante para alguma ação futura. Os verbos que contém a ideia de promessa podem ser enquadrados nesse tipo de ato ilocucionário. A quarta categoria é representada pelos proferimentos expressivos, que indicam sentimentos do falante e são indicadas por verbos como "agradecer", "parabenizar", "desculpar" etc. A quinta e última categoria pertence aos atos declarativos, que trazem alguma alteração no status ou na condição da pessoa ou do objeto referido e podem ser percebidas por frases como "Você está demitido" ou "Eu me demito". A TAF possui pontos controversos em relação à teoria e à classificação dos atos de fala, mas não iremos discuti-los, dados os objetivos desta dissertação, então adotaremos essa abordagem clássica.

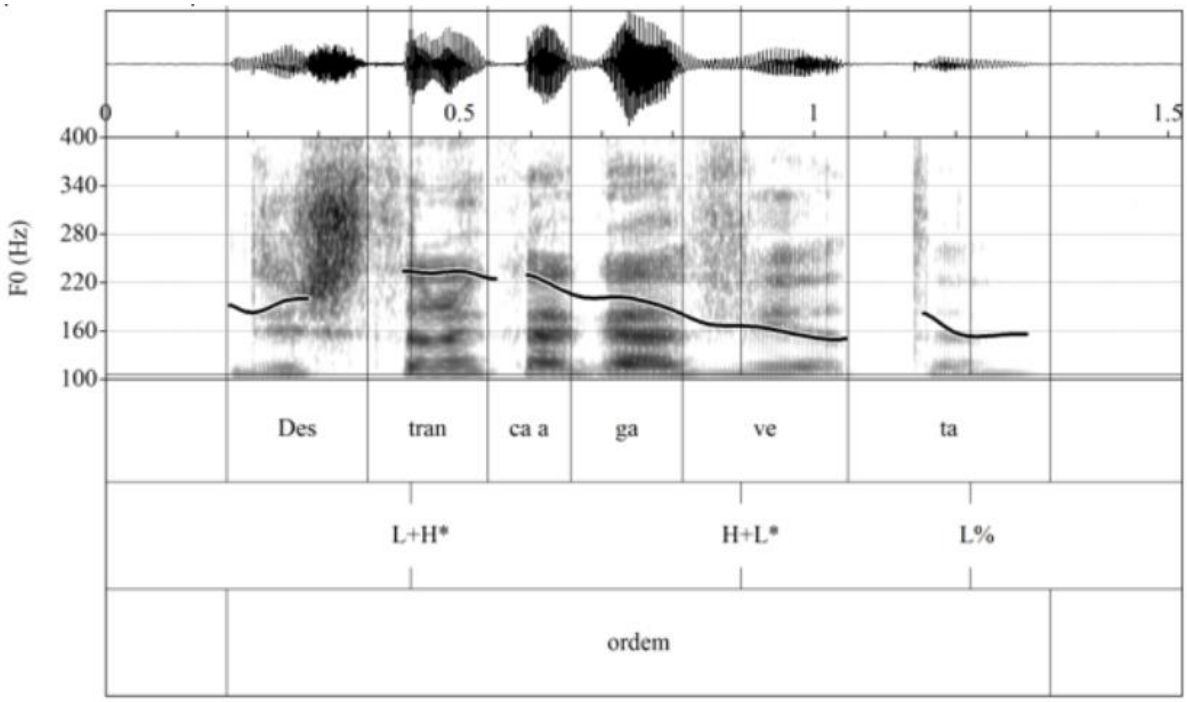
2.3.1 A prosódia da fala diretiva

De acordo com a TAF, a fala diretiva pertence aos atos ilocucionários, ou às ilocuções; o falante usa essas sentenças para induzir seu interlocutor a realizar uma ação, que pode ser exemplificada por atos como uma ordem, um pedido ou uma instrução. Esses atos podem ser realizados prosodicamente de diferentes maneiras, a depender do objetivo do falante (AGUILAR, 2000).

Nos últimos anos, alguns estudos descreveram a prosódia associada ao ato de fala diretivo. A maioria dos trabalhos em português tem como objetivo analisar a entoação da ordem, do pedido e da súplica (QUEIROZ, 2011; BODOLAY, 2014; GOMES DA SILVA *et al.*, 2016; BARRETO, 2019; GOMES DA SILVA *et al.*, 2020) e há outros trabalhos, em menor número, sobre a entoação da sugestão, do desafio (MIRANDA, 2013; MIRANDA & MORAES, 2018) e da instrução (ROCHA, 2011; ROCHA, 2013; RASO & ROCHA, 2015; ROCHA, 2016).

O ato de ordem tem como característica a autoridade que o falante possui diante do interlocutor, e o ato de pedido diferencia-se deste pela possibilidade de recusa (VANDERVEKEN, 1995). Muitas vezes, os atos ilocucionários apresentam a mesma estrutura frasal, tendo a entoação o papel de distinguir essas modalidades (MORAES & COLAMARCO, 2007). Em relação ao contorno melódico, por exemplo, a ordem possui um padrão descendente no português brasileiro, enquanto a súplica contém um padrão circunflexo (MIRANDA & MORAES, 2018; BARRETO, 2019). Na figura 2, é possível ver um exemplo típico da entoação da ordem.

Figura 2 - Contorno melódico da ordem

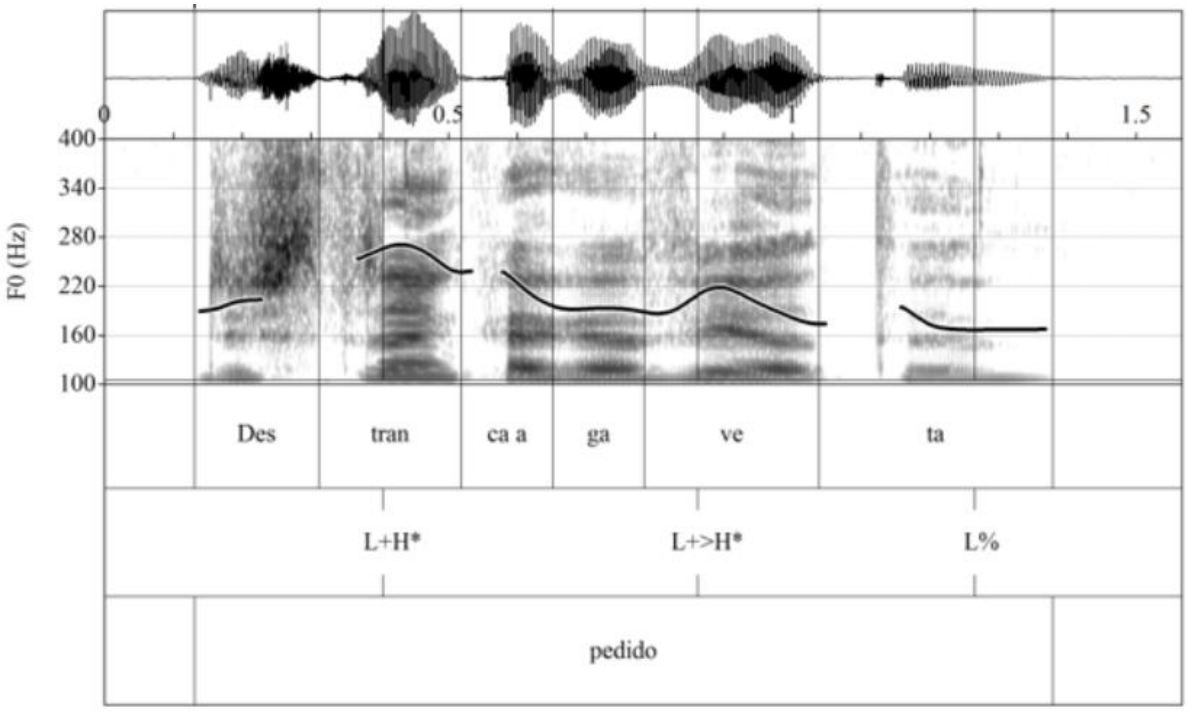


Fonte: Gomes da Silva *et al.* (2020)

Gomes da Silva *et al.* (2020) usam o modelo autossegmental-métrico (AM), que caracteriza os elementos de um sistema de notação entoacional, e explicam que a curva da f0

dos atos de ordem se manifesta com um ataque alto e com uma configuração tonal de L+H* no pré-núcleo e H+L*L% para a posição nuclear. Como se vê, na figura 3, os atos de pedido diferenciam-se deste em posição nuclear com a notação L+>H*L%. De acordo com os autores, não há uma diferença significativa entre os atos de pedido e de súplica no português do Brasil.

Figura 3 - Contorno melódico do pedido



Fonte: Gomes da Silva *et al*. (2020)

As pistas prosódicas podem definir o modo de realização dos diretivos de pedido, de súplica e de ordem, mas os resultados de Queiroz (2011) mostram que esses atos não são categorias estanques, uma vez que há diferentes modos de realização dentro de um mesmo ato diretivo.

Os atos de ordem e de pedido aparecem na maioria dos estudos sobre fala diretiva e prosódia, mas há descrições entoacionais de outros atos, como o desafio e a sugestão. Para Miranda (2013), o desafio possui inicialmente um movimento ascendente e termina com um movimento descendente, enquanto a sugestão também possui um movimento ascendente no início, mas diferencia-se por uma queda contínua final.

Alguns trabalhos que descrevem a prosódia dos atos de fala argumentam que a entoação é um recurso bastante importante no processo de aquisição de uma segunda língua e na formação docente (SANTOS, 2010; BODOLAY, 2020). O que se tem observado é que aprendizes tendem a se basear no sistema da sua língua materna ao proferir sentenças da língua

estrangeira (GOMES DA SILVA, *et al.*, 2011). Para um melhor desempenho dos estudantes estrangeiros, é necessário entender como as pistas prosódicas funcionam na comunicação, a partir de vídeos com fala atuada na internet, de descritores contextuais e de práticas de oralidade, como o *role playing* – encenar alguma situação (BODOLAY, 2020).

Segundo Melo (2017), os atos de ordem e de pedido são pouco explorados em atividades e materiais didáticos no português do Brasil para o aprendizado de língua estrangeira. Não considerar a prosódia da língua-alvo em aulas de língua estrangeira pode ser um problema. Há, por exemplo, diferenças importantes na prosódia dos atos de pedido e de súplica do espanhol, se compararmos essa língua com o português (GOMES DA SILVA *et al.*, 2020).

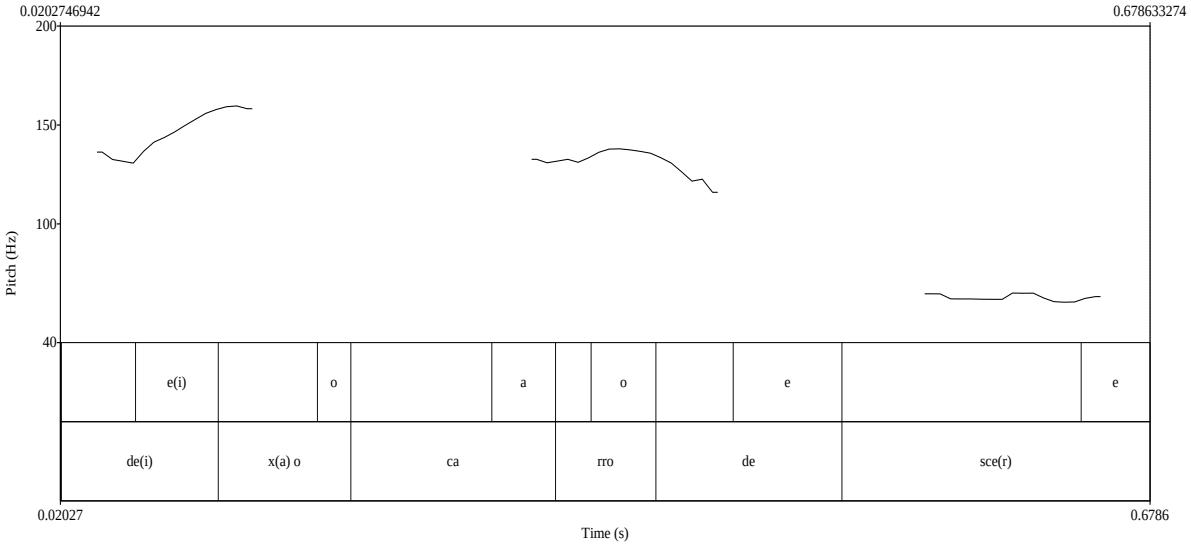
Vê-se, então, que informações prosódicas são muito importantes para o ensino, havendo a necessidade de considerar essa relação quanto à aprendizagem de estudantes. Essa ligação pode ser verificada não apenas nos atos de fala, mas em análises que tentam descrever características prosódicas em segmentações não convencionais de palavras (TENANI; PARANHOS, 2011), como o uso não convencional da vírgula (CARVALHO, 2019). Como mostra Tenani (2017), realizações que fogem do padrão, por exemplo, “em bora” e “senão”, marcadas pela ausência ou não do espaço, em textos do Ensino Fundamental, são resultado de relações morfossintáticas e prosódicas que aparecem na escrita.

2.3.2 A prosódia da instrução

A instrução é um ato diretivo em que o falante guia seu interlocutor, sendo necessário que o instrutor possua os conhecimentos necessários para realizar a ação e que o conteúdo do enunciado contenha expressões linguísticas que indiquem como realizar o procedimento (ROCHA, 2013). A instrução possui um parâmetro pragmático-cognitivo, pois tem como objetivo formar um conhecimento no interlocutor, que deve ter a necessidade de obter esse saber sobre uma ação específica (RASO & ROCHA, 2015; ROCHA, 2016). Uma das diferenças entre a ordem e a instrução explicada por Raso & Rocha (2015) baseia-se no fato de que, na primeira, o falante se beneficia na ação, enquanto que, na instrução, o beneficiário normalmente é o interlocutor.

Rocha (2013) investiga atos de fala relacionados às ilocuções de advertência, oferta, exortação, instrução e ordem. O autor utiliza dados do *corpus* C-ORAL-BRASIL e descreve a prosódia de enunciados de instrução. Um exemplo desse tipo de entoação é mostrado da frase da figura 4.

Figura 4 – Contorno melódico de ilocuções de instrução no *corpus* C-ORAL-BRASIL



Fonte: Rocha (2013)

Conforme Rocha (2013), a frase da figura 4 foi proferida por indivíduos do sexo masculino e tiveram f0 no ataque entre 108 e 134 Hz, f0 mínima entre 80 e 124 Hz, f0 máxima entre 138 e 174 Hz e f0 média entre 112,5 e 141,5 Hz. Rocha (2013) apresenta que a ilocução da ordem contém valores superiores de frequência fundamental se compararmos com a ilocução da instrução.

> A ilocução de Instrução possui um núcleo formado por uma configuração ascendente-descendente de F0 seguida por uma configuração descendente de F0. A configuração ascendente-descendente se inicia em uma das primeiras sílabas do Comentário e pode ser realizada em mais de uma sílaba. A configuração descendente posterior é sempre realizada ao longo das demais sílabas da unidade. O fato da [sic] configuração ascendente-descendente ser seguida por uma configuração descendente na ilocução de Instrução causa a percepção de que se tem uma descida gradual da F0 ao longo da unidade, se comparada à ilocução de Ordem (ROCHA, 2013, p. 135-136).

Rocha (2016) propõe uma metodologia de base pragmática para o estudo das ilocuções, incluindo a instrução, e ressalta que a pesquisa empírica e a pesquisa experimental são necessárias para se ter uma boa descrição da prosódia ilocucionária.

3 METODOLOGIA

Neste capítulo, apresentaremos as fases de nossa pesquisa. Descreveremos o pré-registro e o *corpus* de análise. Em seguida, explicaremos a forma de gravação dos dados e também a segmentação e a anotação. Por fim, delinearemos os procedimentos para a análise dos dados.

3.1 Pré-registro

Esta pesquisa foi pré-registrada[1] no depositório online *Open Science Framework* (OSF), uma plataforma criada pelo *Center for Open Science* (COS) para gerenciar projetos de pesquisa. Segundo Bin & Mota (2022), o pré-registro é a documentação das decisões tomadas a partir da metodologia antes de coletar e analisar os dados, e é a exposição dessa documentação em um repositório online público. Para os autores, é uma prática encorajada pela ciência aberta que visa garantir o rigor e a reprodutibilidade das pesquisas.

3.2 *Corpus* de análise

Este estudo possui uma abordagem descritiva e comparativa, pois busca descrever a entoação da voz do GPS e compará-la com a voz humana. Para atingir esses objetivos, o *corpus* de análise utilizou 20 frases do *Google Maps*, 20 frases do *Waze* e a realização dessas frases por duas participantes nativas do português brasileiro na variedade alagoana, adultas, do sexo feminino e com ensino superior completo. Escolhemos uma voz do sexo feminino, pois o padrão nos aplicativos é a voz feminina.

As frases de um sistema de GPS foram equivalentes às do outro sistema. A escolha desses dois aplicativos (Apps) se deve à sua popularidade em trajetos de carro, em transportes públicos e a pé. Utilizamos a versão 11.28.2 do *Google Maps*, que possui pacotes de voz para vários países, incluindo duas vozes para o Português Brasileiro (PB), uma feminina e outra masculina. Usamos a versão 4.82.5.1 do *Waze*, que possui nove vozes para o PB. Escolhemos uma voz feminina para cada aplicativo, pois é o padrão dos sistemas.

Os aplicativos podem realizar as rotas de modos diferentes e apresentar desempenhos diferentes. Por exemplo, há possibilidade de os números serem descritos isoladamente (como

[1] O link do pré-registro é https://osf.io/57wxn/.

"um, um, cinco" em vez de "cento e quinze") em um dos Apps. Também é possível que um dos aplicativos indique uma direção errada (esquerda ou direita) em relação ao caminho ou use o nome ou o número de uma rodovia de maneira diferente. Por fim, pode haver variação dos verbos utilizados, uma vez que o *Google Maps* tende a ser mais clássico (como "vire à direita"), enquanto o *Waze* pode produzir verbos mais diversificados (como "use a faixa da direita").

3.3 Procedimento da coleta de dados

Para a gravação dos dados de fala humana, utilizamos um gravador digital Marantz da linha PMD661 para registrar as vozes em formato não comprimido WAV, com taxa de amostragem mínima de 96kHz. O microfone utilizado foi o *headset* DPA 4966-TL, devido à sua alta sensibilidade, segundo recomendações de Oliveira Jr. (2014). As gravações foram realizadas em uma cabine acústica do Laboratório de Fonética da Universidade Federal de Alagoas (UFAL). Para garantir a anonimização das participantes e a segurança do procedimento, a pesquisa foi aprovada pelo Comitê de Ética em Pesquisa (CEP)[2] da UFAL. Antes do experimento, as duas participantes leram e assinaram um Termo de Consentimento Livre e Esclarecido (TCLE).

Para a gravação dos dados de fala dos dois softwares de GPS, utilizamos a função de gravação de áudio da Apple, para capturar o som interno do celular. Os dados do *Google Maps* e do *Waze* foram coletados em trechos da cidade de Maceió, de forma que os mesmos estímulos fossem dados aos dois aplicativos. Para isso, foram selecionadas 4 rotas: (I) da rua Hélio Pradines até o Maceió Shopping; (II) da rua Doutor Antônio Gomes de Barros até a rua Deputado José Lages; (III) da rua Deputado José Lages até a Rua Desportista Humberto Guimarães; (IV) da Rua Desportista Humberto Guimarães até a rua Hélio Pradines. As rotas foram pensadas considerando ruas e pontos de referência e incluem curvas, linhas retas e retornos que podem gerar instruções relacionadas a virar, dobrar, entrar, pegar etc., e outras instruções sobre seguir em frente, continuar etc.

Para a gravação dos dados de fala humana, as duas participantes desta pesquisa produziram as mesmas frases realizadas pelos aplicativos *Google Maps* e *Waze*. Uma forma prática de conseguir essa produção é a partir da leitura dessas frases. No entanto, as sentenças do GPS requerem um contexto, uma vez que se trata de falas diretivas.

[2] A avaliação e a aprovação deste estudo foram realizadas pelo Comitê de Ética em Pesquisa da Universidade Federal de Alagoas no dia 3 de dezembro de 2022 (CAAE: 62901922.8.0000.5013).

Para solucionar esse problema, empregamos uma adaptação do instrumento de pesquisa *Discourse Completion Tes*t (DCT), aplicado nos estudos dos atos de fala e já empregado em estudos de atos de fala diretivo, como em Santos (2010), Azuma (2014), Maciel (2015) e Braga *et al.* (2021). O DCT é um teste em que os participantes são submetidos a um conjunto de situações e respondem a um questionário de forma escrita (BILLMYER; VARGHESE, 2000).

Nossa adaptação do DTC propôs que as participantes, separadamente, fossem submetidas a um único contexto e produzissem as frases do *Google Maps* e do *Waze* a partir dele. O comando possui elementos importantes, semelhantes aos de uma conversa comum, como o contexto em que a enunciação ocorre (na rua) e o nível hierárquico entre os participantes (desconhecido): "*Imagine que você está andando em uma rua e um desconhecido lhe pede a localização de um determinado lugar. Com o intuito de ajudar essa pessoa, você profere as seguintes frases:"*. Assim, o teste forneceu informações contextuais claras para as participantes sobre a situação de produção da fala.

Cada frase foi apresentada em um *slideshow* de um dispositivo no momento da aplicação do teste. Cada um dos slides continha uma das frases, que eram passados pelas próprias falantes. As participantes da pesquisa fizeram a leitura do contexto e, em seguida, a produção das frases selecionadas. Assumimos que, como falantes do português, as duas intuitivamente eram capazes de realizar uma entoação de instrução que julgassem adequada. O pesquisador não realizou nenhuma frase com a entoação que ele julgasse adequada, a fim de não influenciar a produção das participantes, mas ressaltou que o objetivo não era imitar a voz do GPS, mas sim produzir as sentenças como se elas estivessem comunicando essas frases a uma determinada pessoa.

É importante mencionar que já existe na literatura uma abordagem metodológica voltada para coletar e analisar as ilocuções no PB (ROCHA, 2013) e para estudar as atitudes dessas ilocuções (ROCHA, 2016). No entanto, a presente pesquisa tem como objeto de estudo a fala do GPS, cujas frases dificilmente são realizadas de forma idêntica em uma conversa espontânea, em razão de seu caráter mais formal. Diante dessa justificativa, os procedimentos desta pesquisa foram realizados conforme descrito acima.

3.4. Seleção das frases

Após as gravações dos dados de fala do GPS, selecionamos as frases para o estudo. Com os mesmos estímulos, o *Google Maps* gerou 44 frases, enquanto o *Waze* gerou 41 frases. Escolhemos 20 frases de cada aplicativo (Apêndice A), pois havia necessidade de padronização

da quantidade, além do fato de que havia repetições de frases em um mesmo App. Incluímos frases equivalentes tanto no *Google Maps* quanto no Waze, e quando não houve mais equivalência, incluímos enunciados conforme a ordem de apresentação até atingir 20 frases. Excluímos frases que não eram diretivas e frases que não atendiam aos critérios de inclusão, por ordem de apresentação (Apêndice B).

Com a gravação das duas falantes, tivemos um total de 80 dados de fala humana, pois a participante 1 leu 20 frases do *Google Maps* e 20 frases do *Waze*, e a participante 2 também. Para comparar a voz humana com as vozes dos aplicativos, selecionamos uma das falas humanas correspondentes a cada uma das frases dos aplicativos. Assim, antes da anotação e da análise acústica, selecionamos 40 dados de fala humana, a fim de compará-los com os 40 dados de fala do GPS. A escolha das frases levou em consideração a qualidade do áudio e a naturalidade que essa frase foi realizada. Além disso, buscando uniformidade na seleção, incluímos vozes das 2 participantes tanto do *Google Maps* quanto do *Waze* (Apêndice C).

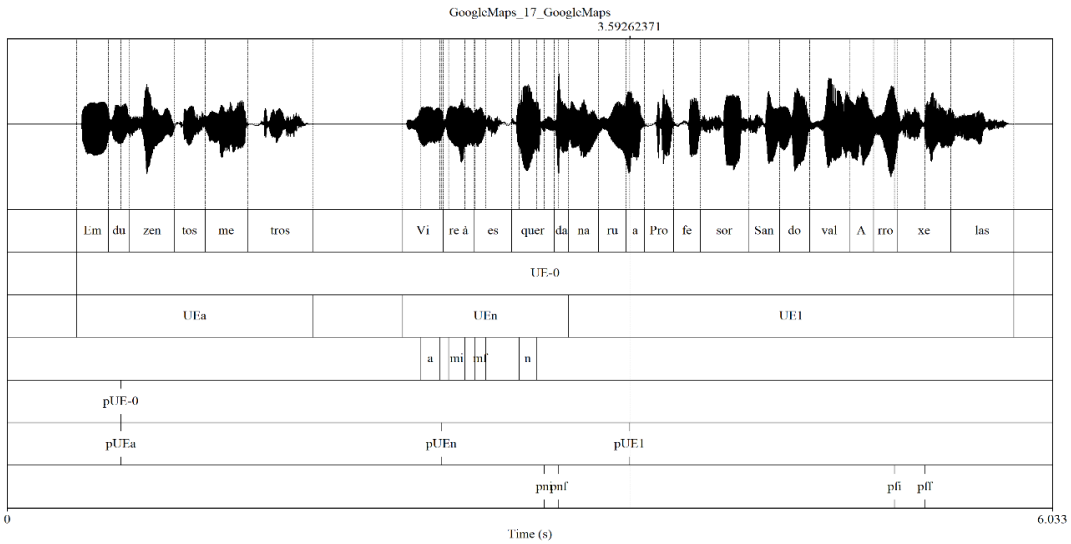
3.5 Segmentação e anotação dos dados

Os dados foram segmentados e anotados a partir do programa Praat. O primeiro passo para uma análise prosódica é identificar a unidade entoacional (UE) (REED, 2011). Esse constituinte prosódico pode ser entendido como um seguimento de palavras agrupadas em um único contorno entoacional (CHAFE, 1988) ou considerado uma unidade linguística formada por aspectos suprassegmentais da fala (REED, 2010).

Entretanto, é importante ressaltar que não há um consenso na literatura sobre a definição de UE (OLIVEIRA JR., 2016). Ao fragmentar a fala em unidades menores, podemos utilizar aspectos prosódicos além da entoação, o que faz com que a UE também seja chamada de unidade prosódica, grupo tonal, unidade terminal, unidade rítmica, frase entoacional ou frase informacional (ALMEIDA, 2017). Apesar disso, adotaremos a expressão unidade entoacional por ser mais conhecida na literatura (BOLINGER, 1989; LIEBERMAN, 1967; OLIVEIRA JR, 2000; BARTH-WEINGARTEN, 2013).

Para delimitar as frases do GPS, usamos uma abordagem intuitiva de segmentação das sentenças, uma vez que falantes nativos concordam significativamente sobre a identificação de unidades entoacionais de uma mesma frase (SILVA & OLIVEIRA JR., 2011; OLIVEIRA JR *et al*., 2012). A anotação dos dados foi realizada de forma multinível, dividida em sete camadas (Figura 5).

Figura 5 - Exemplo de segmentação e anotação das frases



Fonte: elaborado pelo autor (2023)

Na primeira camada (Sílaba), segmentamos e anotamos as frases com uma transcrição ortográfica, sem necessariamente corresponder à divisão silábica da norma culta. Na segunda camada (UE), fragmentamos as frases em unidades entoacionais principais, pois identificamos diferentes UE em uma única enunciação do GPS. Por exemplo, na frase "*Vire à esquerda na Rua Hamilton de Barros Soutinho e depois vire à direita na Rua Soldado Eduardo dos Santos*", há duas unidades principais: "*Vire à esquerda na Rua Hamilton de Barros Soutinho*", anotada como UE-1, e "*e depois vire à direita na Rua Soldado Eduardo dos Santos*", anotada como UE-2. Quando havia apenas uma unidade principal na frase, a sentença era anotada como UE-0.

Também identificamos UEs dentro das unidades principais, segmentadas e anotadas na terceira camada (UE+). Por exemplo, na frase "*Em 200 metros, vire à direita na Rua Abdom Assis Inojosa Andrade*", o trecho "*vire à direita*" corresponde à unidade entoacional nuclear (UEn), que contém o verbo diretivo de ação. A unidade entoacional anterior à UEn foi anotada como UEa ("*Em 200 metros*"). As unidades posteriores à UEn foram anotadas como UE1 e UE2, quando presentes ("*na Rua Abdom Assis Inojosa Andrade*").

Na quarta camada (f0), marcamos o ataque (a), a média inicial (mi), a média final (mf) e o núcleo da sílaba (n) das unidades entoacionais nucleares (UEn). Na quinta camada (pUE), usamos uma marcação de ponto para indicar o primeiro pico de cada uma das unidades principais. Na sexta camada (pUE+), usamos uma camada de ponto para indicar o primeiro pico de cada uma das unidades que formam as unidades principais. Por fim, na última camada (curva3), indicamos os dois últimos picos da unidade entoacional nuclear e os dois últimos picos do final das frases.

3.6 Forma de análise dos dados

A análise acústica, a descrição dos padrões entoacionais e a comparação entre as falas foram realizadas utilizando os *scripts* AnalyseTier (HIRST, 2012) e ProsodyPro (XU, 2013), em conjunto com o Praat. O *script* AnalyseTier forneceu suporte para a análise acústica dos padrões entoacionais, utilizando as camadas geradas durante a segmentação e a anotação dos dados. Também utilizamos o AnalyseTier para obter valores de f0, a fim de calcularmos a taxa de variação e descrever a excursão dos movimentos de f0 a partir dela. Usamos o *script* ProsodyPro, que permite realizar análises sistemáticas de contornos de f0, a partir de pontos-alvo. A fim de comparar o padrão entoacional de dados com diferentes durações, o *script* utiliza um método matemático linear para obter valores de f0 em intervalos regulares (XU, 2013). As curvas de F0 foram comparadas utilizando esse método de alinhamento, o que forneceu um exame dos contornos entoacionais.

Realizamos a análise acústica com o objetivo de investigar se valores de frequência fundamental mínima, média e máxima são diferentes entre as locutoras de aplicativos e humanos. Além disso, analisamos o primeiro pico entoacional das unidades entoacionais e a variação de f0. Partindo da nossa hipótese geral de que a entoação desses dois grupos são diferentes, nossa hipótese era de que esses valores não se aproximariam. Também realizamos uma análise da configuração de f0 e representamos graficamente as curvas entoacionais. Nós supomos que os movimentos da curva seriam diferentes entre aplicativos e humanos e que a representação gráfica das curvas revelaria diferenças entre os dois grupos.

Para validar as comparações entre as falas do GPS e as falas humanas, testes estatísticos foram aplicados por meio do *software* R (R Core Team, 2016-2023), na *interface* RStudio, e as figuras foram geradas neste software a partir do pacote *ggplot2* (WICKHAM, 2016). Utilizamos o teste estatístico de variância (ANOVA) para verificar os níveis de significância entre os níveis de análise, o teste de comparação múltipla de Tukey e o teste t de Welch.

3.6.1. Análise acústica

Nesta subseção, apresentamos os procedimentos metodológicos para as análises acústicas referentes à f0. Os valores correspondentes a esse parâmetro foram extraídos automaticamente com do *script* AnalyseTier, a partir das anotações dos dados.

Adotamos as medidas de f0 mínima, média e máxima como variáveis dependentes para esta análise. Inicialmente, realizamos uma análise da camada UE+, a fim de obter valores das unidades entoacionais formadas pelas unidades entoacionais principais das sentenças (UEa, UEn, UE1, UE2), além de medir a variação de f0, subtraindo os valores de f0 mínima com os de f0 máxima. Depois analisamos os valores do primeiro pico da curva de f0 na camada pUE+ (pUEa, pUEn, pUE1, pUE2).

As análises acústicas tiveram como objetivo verificar o efeito de variáveis nessas duas camadas. A variável independente "locutora" foi formada por falas do *Google Maps* (GM), do *Waze* (WZ), da participante 1 (P1) e da participante 2 (P2). Já a variável independente "voz" foi constituída por TTS (*GoogleMaps* e *Waze*) e HUM (Participante 1 e Participante 2).

No Praat, executamos o AnalyseTier nas camadas mencionadas acima separadamente. Salvamos os resultados em um arquivo *.txt*, abrimos em *.xlsx* em uma planilha do Excel e convertermos para o formato *.csv separado por vírgulas* a fim de automatizar a análise estatística. Os arquivos apresentavam 7 colunas:

(1) *source* – nome do aplicativo referente à frase (GoogleMaps ou Waze);

(2) *num* – número da frase de cada um dos Apps (1, 2... 20);

(3) *speaker* – variável locutora;

(3) *voice* – variável voz;

(4) *UE+* – (UEa, UEn, UE1, UE2); *pUE+* – (pUEa, pUEn, pUE1, pUE2), isto é, marcações das camadas separadamente em arquivos *.csv*;

(5) *f0_min* – frequência fundamental mínima;

(6) *f0_mean* – frequência fundamental média;

(7) *f0_max* – frequência fundamental máxima;

(8) *variação de f0* – valores da subtração (= *f0_max – f0_min*).

3.6.2 Análise de padrão da curva entoacional

Para analisar o padrão da curva entoacional, adaptamos um método de Rocha (2016) para verificar a taxa de variação melódica e a excursão dos movimentos de f0. Esse procedimento foi necessário para verificar a ocorrência de movimentos ascendentes ou descendentes. Para Rocha (2016), a taxa de variação (*(f0 final – f0 inicial)* / tempo) reflete a inclinação de um movimento de f0. Para medir a f0 final e f0 inicial neste trabalho, utilizamos os valores gerados pela camada f0.

Consideramos os valores de f0 máxima do ataque (a), da média inicial (mi) e da média final (mf) da camada f0, bem como os valores dos últimos picos entoacionais anotados na camada curva3 para verificar a taxa de variação e a excursão dos movimentos. Usamos a f0 máxima (ou *pitch range*)[3] como medida de variação, pois é considerada como ideal para esse fim (LIBERMAN & PIERREHUMBERT, 1984; OLIVEIRA JR., 2000; SWERTS, 1997).

Para esta análise, verificamos as unidades entoacionais nucleares (UEn), pois é onde está localizado o verbo de ação da frase diretiva, além de a extensão do trecho ser pequena o suficiente para ser comparada com as frases já descritas pela literatura. Analisamos também a curva final de toda a frase.

Descrevemos três movimentos na curva, a fim de comparar com os padrões entoacionais típicos da instrução. Para indicação do primeiro movimento, subtraímos os valores de (mi) com os de (a). Para o segundo movimento, subtraímos os valores de (mf) com os valores de (mi). Em seguida, calculamos as médias desses resultados no R. Se o número fosse negativo, ou seja, abaixo de 0, isso indicava um contorno descendente; se o resultado fosse positivo, o movimento seria ascendente (ROCHA, 2016).

Para a indicação do terceiro movimento, foi necessário usar a camada curva3, que indicava os últimos picos da unidade entoacional nuclear e os últimos picos do final de toda frase. Isso foi necessário, pois havia dois contextos de realização nesse movimento. No primeiro contexto, a maioria das UEn do GPS aparecia no início ou no meio das frases (por exemplo: *Em 200 metros, **vire à esquerda** na rua Professor Sandoval Arroxelas*). No segundo contexto, as UEn apareciam no fim da frase (por exemplo: *Em 300 metros, **vire à esquerda***), em menor número. Assim, para análise desse terceiro movimento, subtraímos os valores do último pico com os do penúltimo pico no primeiro contexto e no segundo contexto de UEn.

A taxa de variação e a excursão de movimento foram medidas tanto em Hertz e quanto em semitons, separadamente. Isso permitiu observar se havia diferença no nível da produção (f0) e da percepção (*pitch*) da fala.

[3] Em geral, a f0 máxima é considerada como corresponde ao *pitch range* (LIBERMAN & PIERREHUMBERT, 1984; SWERTS, 1997; OLIVEIRA JR., 2000).

3.6.3 Representação da curva entoacional

Usamos o *script* ProsodyPro e o software R para obter uma representação gráfica da curva entoacional das unidades entoacionais nucleares (UEn). O *script* é capaz de identificar pontos-alvos de f0 em uma determinada camada segmentada e anotada no Praat.

Inicialmente, selecionamos as frases em que as UEn apareciam no início ou no meio e as frases em que as UEn apareciam no final das sentenças. Em seguida, separamos o áudio e o *textgrid* relativos às duas participantes (HUM) dos arquivos dos dois aplicativos de GPS (TTS). Rodamos o *script* a partir da camada UE+, identificando 10 pontos-alvo de f0.

Em uma mesma frase, havia UEn no início/meio e UEn que estavam no fim, além dos outros tipos de unidades entoacionais. Então, tomamos medidas cuidadosas para garantir que todas UEn fossem criteriosamente selecionadas e separadas. Os valores foram exportados para uma planilha e calculamos as médias de TTS e de HUM nos dois contextos de UEn, utilizando a função do Excel. Por fim, geramos um gráfico, usando o pacote *ggplot2* do R.

4 ANÁLISE DOS DADOS

Esta seção apresenta os resultados da análise dos dados em relação à análise acústica, à descrição entoacional e à comparação dos contornos. Durante essa exposição, retomamos a descrição das abreviações e das camadas, pois acreditamos que isso facilita a compreensão.

4.1 Análise acústica da camada UE+

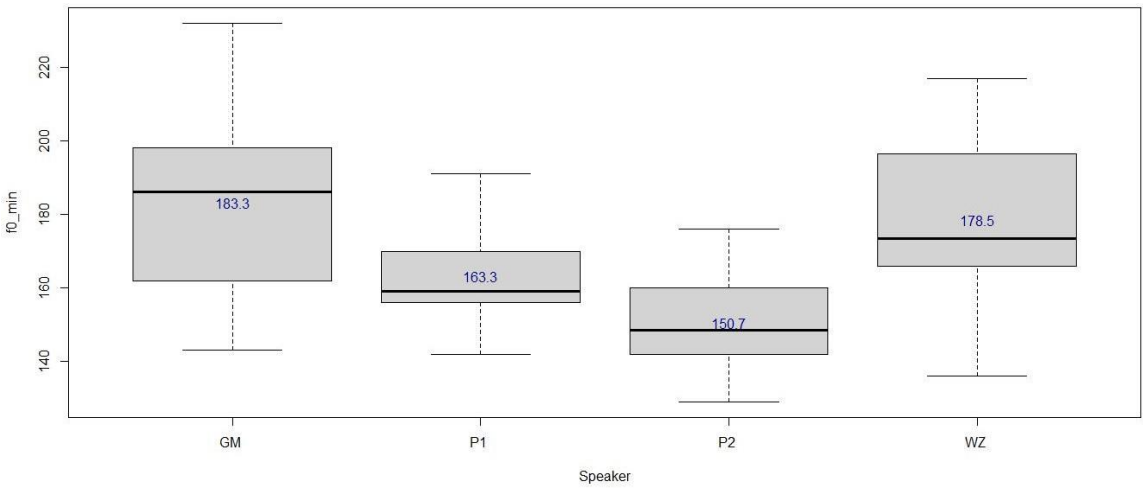
A camada UE+ representa as anotações da unidade entoacional anterior à unidade nuclear (UEa), da unidade entoacional nuclear (UEn), em que está o verbo de ação, da primeira unidade entoacional depois da nuclear (UE1) e da segunda unidade entoacional depois da nuclear (UE2).

Nas subseções a seguir, descreveremos os resultados da análise acústica da frequência fundamental nessa camada. Essa análise apresenta informações sobre a variável locutora – *Google Maps* (GM), *Waze* (WZ), participante 1 (P1) e participante 2 (P2) – e a variável tipo de voz – aplicativos (TTS) e humanos (HUM).

4.1.1 Efeito da variável locutora nas médias de f0_min da camada UE+

Um teste ANOVA foi realizado para comparar o efeito das locutoras na frequência fundamental mínima (f0_min). O teste revelou uma diferença significativa entre pelo menos duas locutoras ($F (3, 226) = 44.77$, $p < 2e\text{-}16$) (Gráfico 1).

Gráfico 1 - Boxplot f0_min por locutora da UE+



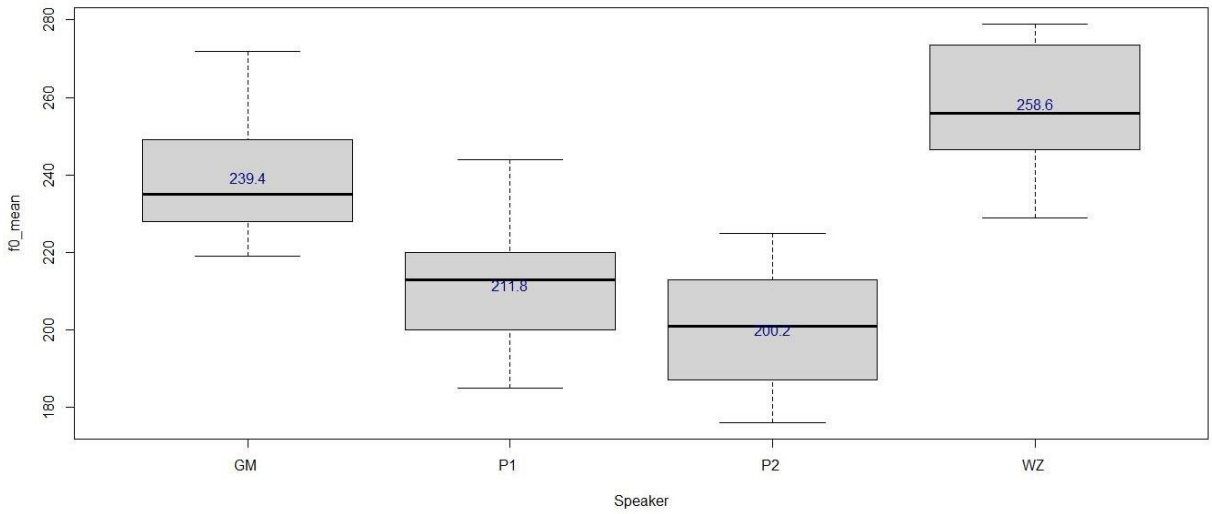
Fonte: elaborado pelo autor (2023)

O teste de comparações múltiplas de Tukey foi realizado para comparar as médias de f0_min entre as locutoras, a fim de verificar quais foram os grupos que não houve diferença. Os resultados mostraram diferenças estatisticamente significativas entre as locutoras P1 e GM (p = 0.0000000), P2 e GM (p = 0.0000000), P2 e P1 (p = 0.0007072), WZ e P1 (p = 0.0000356) e WZ e P2 (p = 0.0000000). Não houve diferença estatisticamente significativa entre WZ e GM (p = 0.4427198).

4.1.2 Efeito da variável locutora nas médias de f0_mean da camada UE+

Realizamos também um teste ANOVA para verificar o efeito da variável locutora na variável dependente f0_mean (frequência fundamental média). O resultado do teste mostrou que o efeito do fator locutora foi significativo (F(3,226) = 11, p <2e-16), indicando que existe uma diferença entre as diferentes locutoras em relação à f0 média (Gráfico 2).

Gráfico 2 - Boxplot f0_mean por locutora da UE+



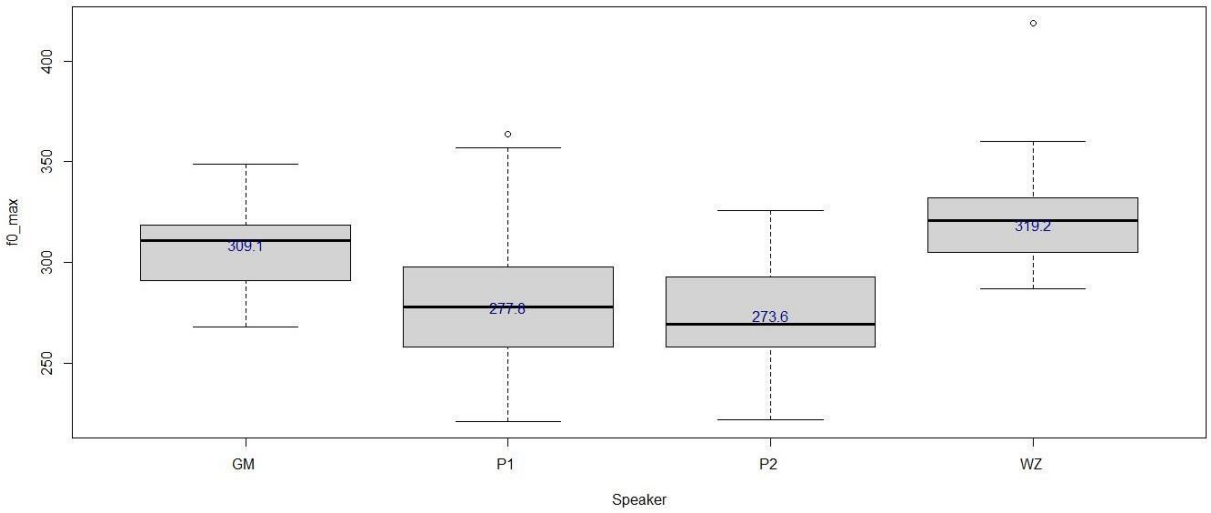
Fonte: elaborado pelo autor (2023)

Para entender melhor como esses grupos diferem entre si, realizamos um teste *post-hoc* de comparações múltiplas. O teste indicou que as médias entre todas as locutoras são diferentes. A média da participante 1 foi significativamente menor que a do *Google Maps* (diferença = -27.61625, p < 0.001), assim como a média da participante 2 foi significativamente menor que a do *Google Maps* (diferença = -39.21241, p < 0.001). Por outro lado, a média do *Waze* foi significativamente maior que a do *Google Maps* (diferença = 19.16374, p < 0.001).

4.1.3 Efeito da variável locutora nas médias de f0_max da camada UE+

Realizamos um teste ANOVA para verificar o efeito da variável locutora na variável depende f0_max (frequência fundamental máxima). O resultado mostrou que há uma diferença estatisticamente significativa entre as médias de f0 máxima de pelo menos duas locutoras (F (3, 226) = 52.92, p < 2e-16) (Gráfico 3).

Gráfico 3 - Boxplot f0_max por locutora da UE+



Fonte: elaborado pelo autor (2023)

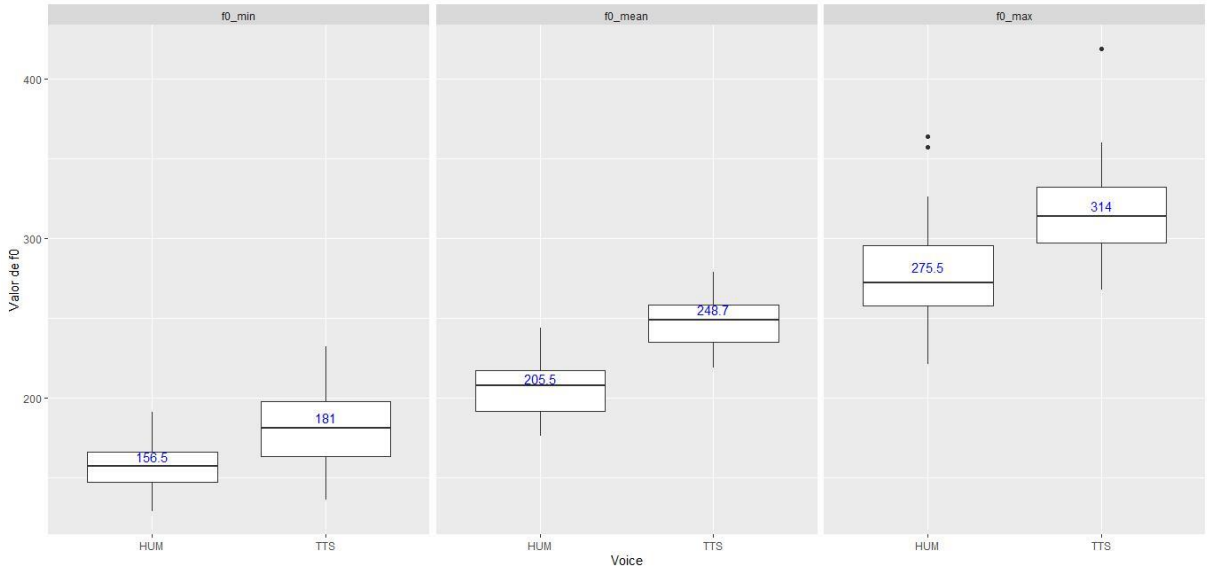
Usamos um teste *post-hoc* para determinar quais locutoras apresentam diferenças significativas nas médias. O teste de comparações múltiplas de Tukey foi realizado para comparar as médias de f0_max entre os grupos. Os resultados mostraram que houve uma diferença estatisticamente significativa entre P1 e GM (p = 0.0000000), P2 e GM (p = 0.0000000), WZ e P1 (p = 0.0000000) e WZ e P2 (p = 0.0000000).

Não houve diferença estatisticamente significativa entre as falas dos aplicativos (p = 0.1058827) ou entre as duas participantes (p = 0.7776306). Como a f0_max é o correlato ao *pitch range*, que é uma medida que afeta a escala geral do contorno de f0 (LIBERMAN & PIERREHUMBERT, 1984), esses resultados apontaram para uma distinção clara entre os grupos HUM e TTS.

4.1.4 Teste t para efeito da variável voz nas médias da camada UE+

Realizamos o teste t de duas amostras de Welch para comparar o efeito do tipo de voz (TTS e HUM) nas médias de f0_min, f0_mean e f0_max. As estimativas de amostra das médias mostraram diferenças entre os dois grupos (Gráfico 4).

Gráfico 4 - Boxplot tipo de voz nas médias de f0_min, f_mean e f0_max da UE+



Fonte: elaborado pelo autor (2023)

O teste t revelou que houve diferenças estatisticamente significativas na f0_min entre os dois grupos (t (188.46) = -10.458, p < 2.2e-16), na f0_mean entre os dois grupos (t (225.68) = -20.567, p < 2.2e-16) e na f0_max entre os dois grupos (t (218.94) = -12.246, p < 2.2e-16).

4.1.5 Efeito das variáveis voz e UE nas médias de f0_min da camada UE+

Realizamos um teste ANOVA de duas vias com interação para avaliar o efeito do tipo de voz (TTS e HUM) e UE na variável dependente f0_min. Os resultados mostraram que houve um efeito principal significativo da voz (F(1, 222) = 136.057, p < 0.001) e da UE (F(3, 222) = 18.181, p < 0.001) na variável dependente. No entanto, a interação entre voz e UE não foi significativa (F(3, 222) = 2.356, p = 0.0728).

Um teste de comparações múltiplas de Tukey também foi realizado para avaliar diferenças entre os níveis das variáveis voz e UE. Os resultados indicaram uma diferença significativa entre os níveis TTS e HUM da variável voz, com uma diferença média de 24.47826 (Intervalo de Confiança 95%: [20.34262, 28.6139]), p < 0.001. Para a variável UE, entre

humanos e aplicativos, houve diferenças significativas nas UEn e UE1, mas não houve nas UEa e UE2 (Tabela 1).

Tabela 1 - Valor *p* em relação ao teste de interação entre voz e UE na f0_min

Tipo de voz: contexto	Tipo de voz: contexto	Valor *p*
TTS: UEa	HUM: UEa	0.1784214
TTS: UEn	HUM: UEn	0.0000000
TTS: UE1	HUM: UE1	0.0000001
TTS: UE2	HUM: UE2	0.5907290

Fonte: elaborado pelo autor (2023)

4.1.6 Efeito das variáveis voz e UE nas médias de f0_mean da camada UE+

Um teste de interação ANOVA de dois fatores foi realizado para avaliar o efeito de voz e UE na variável dependente f0_mean. Os resultados indicaram um efeito significativo de voz $F(1, 222) = 599.915$, $p < 0.001$ e de UE na f0_mean, $F(3, 222) = 33.086$, $p < 0.001$. No entanto, a interação entre voz e UE não foi significativa, $F(3, 222) = 0.699$, $p = 0.553$.

Realizamos também um teste de comparações múltiplas de Tukey para avaliar diferenças entre os níveis das variáveis voz e UE. Os resultados indicaram uma diferença significativa entre os níveis TTS e HUM da variável voz, com uma diferença média de 43.2 (IC 95%: [39.72415, 46.67585]), $p < 0.001$. Para a variável UE, houve diferenças significativas nas UEa, UEn, UE1 e UE2 (Tabela 2).

Tabela 2 - Valor *p* em relação ao teste de interação entre voz e UE na f0_mean

(Tipo de voz: contexto)	(Tipo de voz: contexto)	Valor *p*
TTS: UEa	HUM: UEa	0.0000000
TTS: UEn	HUM: UEn	0.0000000
TTS: UE1	HUM: UE1	0.0000000
TTS: UE2	HUM: UE2	0.0241937

Fonte: elaborado pelo autor (2023)

4.1.7 Efeito das variáveis voz e UE nas médias de f0_max da camada UE+

Um teste de interação ANOVA de dois fatores foi realizado para avaliar o efeito da voz e da UE na variável dependente f0_max. Os resultados indicaram efeitos significativos de voz, $F(1, 222) = 170.793$, $p < 0.001$, e de UE na f0_max, $F(3, 222) = 9.633$, $p < 0.01$. Além disso, a interação entre voz e UE foi significativa, $F(3, 222) = 2.924$, $p < 0.05$.

Para avaliar diferenças entre os níveis das variáveis voz e UE, foi realizado um teste *post-hoc* de comparações múltiplas. Os resultados indicaram uma diferença significativa entre os níveis TTS e HUM da variável voz, com uma diferença média de 38.47826 (IC 95%: [32.67593, 44.2806]), $p < 0.001$. Para a variável UE, houve diferenças significativas nas UEa, UEn, UE1 e UE2 (Tabela 3).

Tabela 3 - Valor de p em relação ao teste de interação entre voz e UE na f0_max

(Tipo de voz: contexto)	(Tipo de voz: contexto)	Valor *p*
TTS: UEa	HUM: UEa	0.0000064
TTS: UEn	HUM: UEn	0.0000000
TTS: UE1	HUM: UE1	0.0000034
TTS: UE2	HUM: UE2	0.0303489

Fonte: elaborado pelo autor (2023)

4.1.8 Variação de f0

Medimos a variação de f0, subtraindo os valores de f0_max com os valores de f0_min na camada UE+. As médias da subtração no grupo TTS foi de (133,0522), enquanto no grupo HUM foi de (119,0522).

Realizamos um teste t de Welch para comparar essas médias. O teste indicou uma diferença significativa nos dois grupos ($t(227,46) = -4,1401$, $p < 4.892e-05$), com intervalo de confiança de 95% para a diferença entre as médias de 20,66 a 7,34.

4.2. Análise acústica da camada pUE+

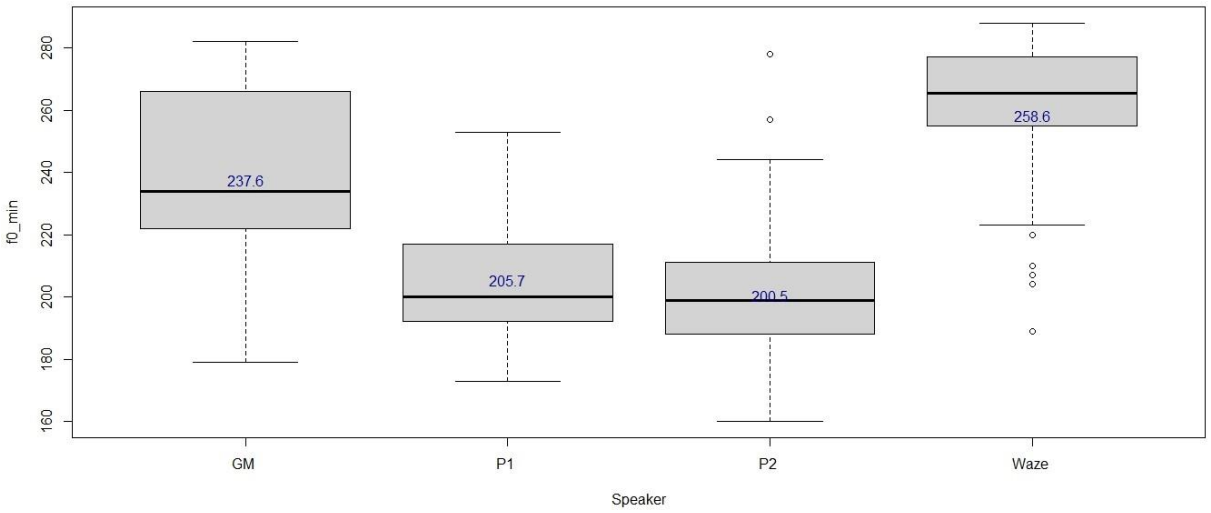
A camada pUE+ indica as anotações do primeiro pico de cada unidade entoacional das unidades entoacionais principais, ou seja, o primeiro pico entoacional da unidade anterior à unidade nuclear (pUEa), da unidade entoacional nuclear (pUEn), em que está o verbo de ação, da primeira unidade entoacional depois da nuclear (pUE1) e da segunda unidade entoacional depois da nuclear (pUE2).

Nas subseções a seguir, apresentaremos os resultados da análise acústica da frequência fundamental nessa camada. Esses achados mostram informações sobre o primeiro pico entoacional realizados a partir da variável locutora e da variável tipo de voz.

4.2.1 Efeito da variável locutora nas médias de f0_min da camada pUE+

Realizamos um teste ANOVA para o fator "locutora" na frequência fundamental mínima (f0_min) da camada pUE+. O teste revelou um resultado F (3, 226) = 69.35, p < 2e-16, indicando que as diferenças entre as médias dos grupos são estatisticamente significativas (Gráfico 5).

Gráfico 5 - Boxplot f0_min por locutora da pUE+



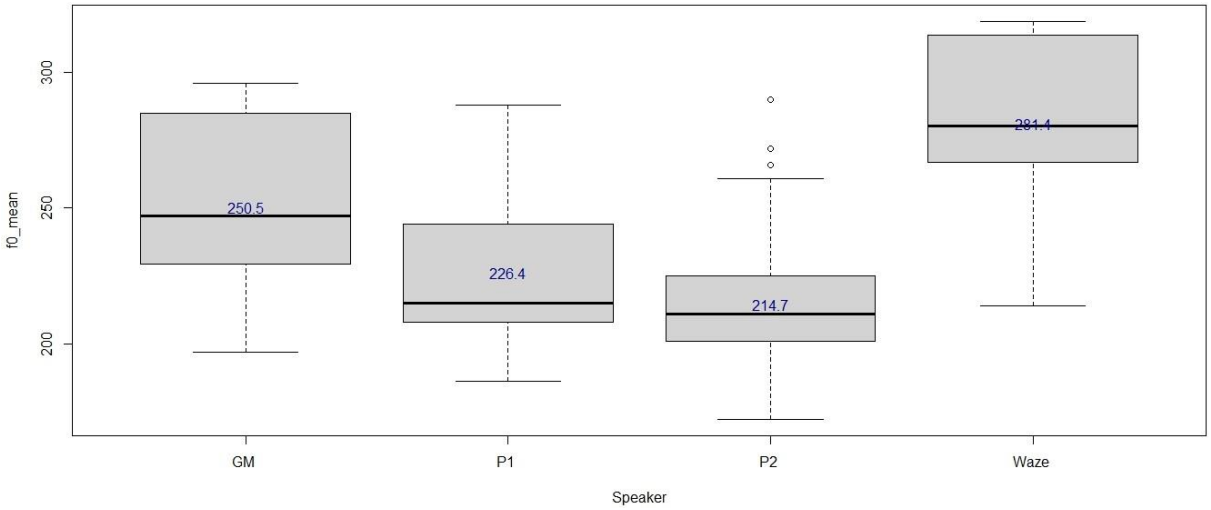
Fonte: elaborado pelo autor (2023)

Usamos um teste de Tukey para comparar as médias das locutoras em relação à f0_min. Os resultados mostraram que houve diferença estatisticamente significativa entre a participante 1 e o *Google Maps* (p = 0.0000000), a participante 2 e o *Google Maps* (p = 0.0000000), as locutoras dos dois aplicativos (p = 0.0000631), o *Waze* e a participante 1 (p = 0.0000000) e o *Waze* e a participante 2 (p = 0.0000000). Não houve diferença significativa apenas entre o primeiro pico entoacional da participante 1 e o da participante 2 na f0_min (p = 0.6818628).

4.2.2 Efeito da variável locutora nas médias de f0_mean da camada pUE+

Fizemos também um teste ANOVA para o fator locutora na variável dependente f0_mean (frequência fundamental média). O teste sugeriu que as diferenças entre as médias dos grupos são estatisticamente significativas (F (3, 226) = 63.83, p < 2e-16) (Gráfico 6).

Gráfico 6 - Boxplot f0_mean por locutora da pUE+



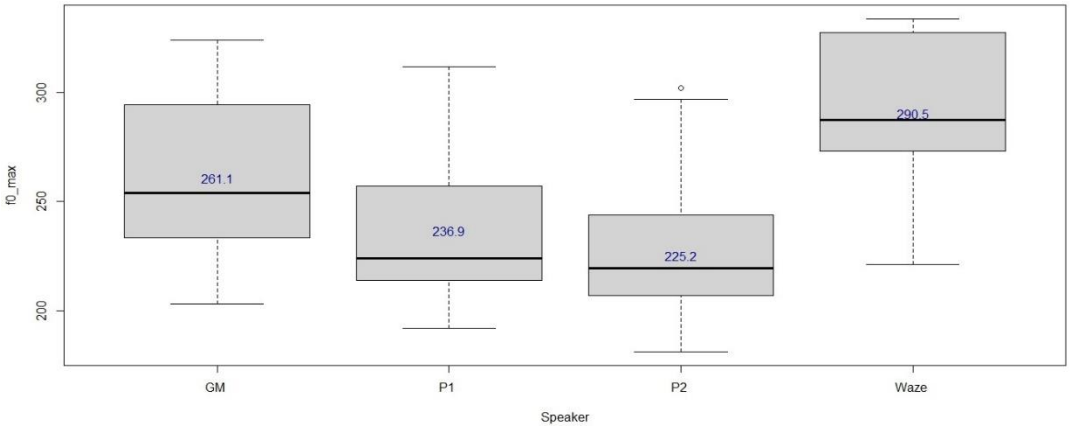
Fonte: elaborado pelo autor (2023)

A fim de entender melhor como as quatro locutoras diferem entre si, realizamos um teste *post-hoc* de comparações múltiplas. O resultado obtido foi que não houve diferença significativa apenas entre as duas locutoras humanas (P1 – P2, *p* = 0.1170745), enquanto houve diferença entre as demais locutoras (P1 – GM, *p* = 0.0000538; P2 – GM, *p* = 0.0000000; WZ – GM, *p* = 0.0000001; WZ – P1, *p* = 0.0000000, WZ – P2, *p* = 0.0000000).

4.2.3 Efeito da variável locutora nas médias de f0_max da camada pUE+

Fizemos um teste ANOVA para verificar o efeito da variável locutora na variável depende f0_max da camada referente ao primeiro pico entoacional das UE. O teste revelou que as diferenças entre as médias dos grupos são estatisticamente significativas, exceto entre dois grupos (F (3, 226) = 46.44, p < 2e-16) (Gráfico 7).

Gráfico 7 - Boxplot f0_max por locutora da pUE+



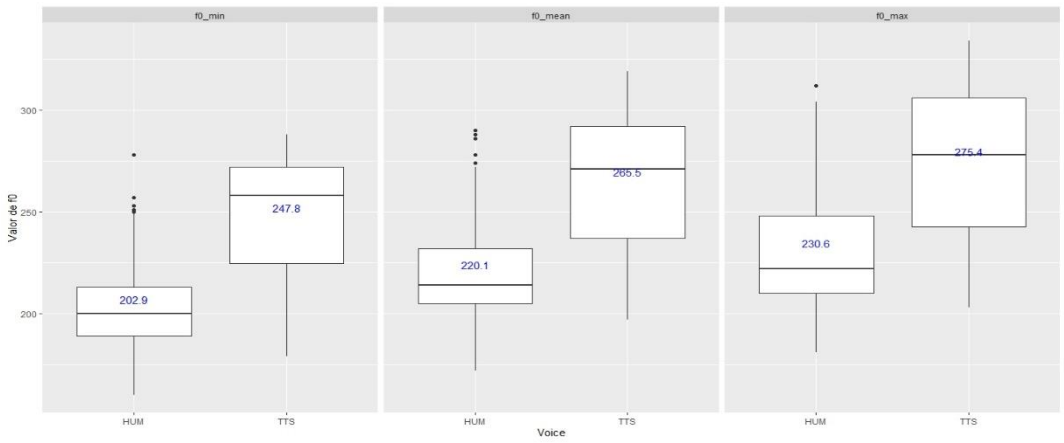
Fonte: elaborado pelo autor (2023)

Utilizamos um teste de Tukey *post-hoc* para comparar as médias de f0_max entre os grupos de locutoras. Os dois grupos que não apresentaram diferença significativa entre eles foram as locutoras P1 e P2 (p = 0.2115857), enquanto houve diferença entre os demais grupos (P1 – GM, p = 0.0005528; P2 – GM, p = 0.0000000; WZ – GM, p = 0.0000110; WZ – P1, p = 0.0000000; WZ – P2, p = 0.0000000).

4.2.4 Teste t para efeito da variável voz nas médias da camada pUE+

Usamos o teste t de duas amostras de Welch para comparar o efeito do tipo de voz (TTS e HUM) nas médias de f0_min, f0_mean e f0_max no primeiro pico entoacional das UE. As estimativas de amostra das médias mostraram diferenças entre os dois grupos (Gráfico 8).

Gráfico 8 - Boxplot tipo de voz nas médias de f0_min, f_mean e f0_max da pUE+



Fonte: elaborado pelo autor (2023)

O valor t obtido nas médias de f0_min foi de -13.108 com valor *p* menor que 2.2e-16, enquanto que, nas médias de f0_mean, foi de -11.391 com valor *p* menor que 2.2e-16, e nas médias de f0_max foi de -10.011 com valor *p* menor que 2.2e-16. A partir desse teste, é possível verificar que, independentemente da medida de f0 analisada, a diferença entre as médias de aplicativos e de humanos nessa camada é estatisticamente significativa.

4.2.5 Efeito das variáveis voz e UE nas médias de f0_min da camada pUE+

Utilizamos um teste ANOVA de dois fatores para verificar o efeito do tipo de voz e de pUE, e a interação entre eles na variável dependente f0_min. O valor F obtido para o fator voz foi de 250.603, e o valor *p* foi menor que 2e-16, indicando que as diferenças entre as médias dos grupos são estatisticamente significativas. O valor F obtido para o fator pUE foi de 34.178 e o valor *p* também foi menor que 2e-16, indicando diferenças entre as médias dos grupos. O valor F obtido para a interação entre os fatores tipo de voz e pUE foi de 2.665, e o valor *p* foi de 0.0487, indicando que a interação é estatisticamente significativa. Em outras palavras, isso sugere que o efeito do fator voz na variável f0_min depende do nível do fator pUE.

Para a variável voz, o teste de Tukey mostrou diferença entre as médias dos grupos TTS e HUM de 44.94, com um intervalo de confiança de 95% entre 39.34 e 50.53, com um valor *p* ajustado igual a 0, indicando que a diferença é estatisticamente significativa. Para a variável pUE, houve diferenças significativas em três contextos, exceto no pUE2 (Tabela 4).

Tabela 4 - Valor *p* em relação ao teste de interação entre voz e pUE na f0_min

Tipo de voz: contexto	Tipo de voz: contexto	Valor *p*
TTS: pUEa	HUM: pUEa	0.0000002
TTS: pUEn	HUM: pUEn	0.0000000
TTS: pUE1	HUM: pUE1	0.0000000
TTS: pUE2	HUM: pUE2	0.9907387

Fonte: elaborado pelo autor (2023)

4.2.6 Efeito das variáveis voz e UE nas médias de f0_mean da camada pUE+

Foi realizado um teste ANOVA de dois fatores para avaliar o efeito dos fatores tipo de voz e pUE e da interação entre eles na variável f0_mean. O valor F obtido para o fator voz foi

de 222.684 com valor *p* menor que 2e-16, enquanto o valor F obtido para o fator pUE foi de 53.897 com valor *p* também menor que 2e-16, mostrando que as diferenças entre as médias dos grupos são estatisticamente significativas. O valor F obtido para a interação entre os fatores voz e pUE foi de 2.542 com valor *p* de 0.0572, indicando que a interação não é estatisticamente significativa ao nível de significância de 0.05.

Também foi realizado um teste de Tukey para comparar as médias dos grupos voz e pUE e suas interações. Para o fator voz, a diferença entre as médias dos grupos TTS e HUM foi de 45.48, com um intervalo de confiança de 95% entre 39.47 e 51.48 e um valor *p* ajustado igual a 0, indicando que a diferença é estatisticamente significativa. Para a variável pUE, houve diferenças significativas em pUEa, pUEn, pUE1, mas não houve diferenças em pUE2 (Tabela 5).

Tabela 5 - Valor *p* em relação ao teste de interação entre voz e pUE na f0_mean

Tipo de voz: contexto	Tipo de voz: contexto	Valor *p*
TTS: UEa	HUM: UEa	0.0000001
TTS: UEn	HUM: UEn	0.0000000
TTS: UE1	HUM: UE1	0.0000000
TTS: UE2	HUM: UE2	0.9748655

Fonte: elaborado pelo autor (2023)

4.2.7 Efeito das variáveis voz e UE nas médias de f0_max da camada pUE+

O resultado do teste ANOVA de dois fatores para avaliar o efeito de voz e de pUE na variável f0_max indicou efeitos significativos de voz (valor F de 179.148 com valor *p* menor que 2e-16) e de pUE (valor F de 60.260 com valor *p* também menor que 2e-16). O valor F obtido para a interação entre os fatores voz e pUE foi de 1.587, e o valor *p* foi de 0.193, sugerindo que a interação não é estatisticamente significativa ao nível de significância de 0.05.

A partir do teste de Tukey, para a variável voz, a diferença entre as médias dos grupos TTS e HUM foi de 44.8, com um intervalo de confiança de 95% entre 38.20 e 51.40, com um valor *p* ajustado igual a 0, indicando que a diferença é estatisticamente significativa. Para a variável pUE, houve diferenças significativas em todos os contextos, exceto no pUE2 (Tabela 6).

Tabela 6 - Valor *p* em relação ao teste de interação entre voz e pUE na f0_max

Tipo de voz: contexto	Tipo de voz: contexto	Valor *p*
TTS: UEa	HUM: UEa	0.0000006
TTS: UEn	HUM: UEn	0.0000000
TTS: UE1	HUM: UE1	0.0000000
TTS: UE2	HUM: UE2	0.9235037

Fonte: elaborado pelo autor (2023)

4.3 Taxa de variação e excursão dos movimentos da curva entoacional

Apresentamos, nesta subseção, os resultados referentes à taxa de variação e à excursão dos movimentos da curva entoacional das UEn. Para isso, utilizamos os valores obtidos do AnalyseTier, em hertz e em semitons, a partir da camada f0 da anotação dos dados, com os valores de f0 de ataque (a), média inicial (mi) e média final (mf). Utilizamos também a camada curva3 a fim de verificar a entoação do terceiro movimento.

4.3.1 Taxa de variação e excursão do primeiro movimento da curva entoacional

Executamos o AnalyseTier na camada f0 e medimos os valores em hertz. Após subtrair os resultados de (mi) com os de (a) em cada frase analisada, calculamos a média das diferenças, que foi de (10, 87037) no grupo TTS e (12, 37037) no grupo HUM. O teste t de Welch mostrou que não houve diferença significativa entre as médias dos dois grupos (p = 0.5746).

Utilizamos também os valores em semitons no AnalyseTier. A média das diferenças nos aplicativos foi de (282, 1730), e a dos humanos foi de (41, 6524). Os resultados referentes ao teste t sugerem que houve uma diferença significativa entre os dois grupos (p = 0.001483). Avaliando os resultados como um todo e individualmente, é possível perceber um primeiro movimento ascendente na curva da UEn, tanto nos aplicativos quanto nos humanos.

4.3.2 Taxa de variação e excursão do segundo movimento da curva entoacional

Após medirmos os valores em hertz e calcularmos as diferenças entre (mf) e (mi), o grupo dos aplicativos mostrou uma média de (-44,46296), enquanto o dos humanos apresentou (-6,433962). O teste t revelou uma diferença significativa entre as médias dos dois grupos com valor p = 4.429e-09.

Os valores em semitons e o cálculo das diferenças entre (mf) e (mi) apresentaram uma média de (-112,3537) para TTS e (-72,8880) para HUM. Não houve uma diferença significativa entre as médias dos grupos conforme o teste t (p = 0.5798). Os resultados em hertz e em semitons indicaram uma configuração descendente no segundo movimento da curva entoacional das UEn em ambos os grupos.

4.3.3 Taxa de variação e excursão do terceiro movimento da curva entoacional

Para a análise do terceiro movimento, utilizamos os valores obtidos pelo AnalyseTier na camada curva3 das UEn e do final das frases do GPS. Depois de subtrair os valores do último pico com os do penúltimo pico, as médias em hertz mostraram um movimento ascendente nas nucleares que aparecem no início ou no meio da frase e um movimento descendente no final das frases do GPS (Tabela 7).

Tabela 7 - Taxa de variação do terceiro movimento da curva em hertz

Contexto	Tipo de voz	Taxa de variação
UEn no início ou no meio da frase	TTS	24,74419
UEn no início ou no meio da frase	HUM	31,11628
Final das frases do GPS	TTS	-73,02500
Final das frases do GPS	HUM	-37,61538

Fonte: elaborado pelo autor (2023)

Os testes t revelaram que não houve diferenças significativas entre as médias em hertz dos dois grupos nas UEn que aparecem no início/meio da frase (p = 0.2627), mas houve diferença nas frases finais do GPS (p = 1.017e-05).

As médias em semitons também mostraram um movimento ascendente nas UEn e um movimento descendente no final das frases (Tabela 8).

Tabela 8 - Taxa de variação do terceiro movimento da curva em semitons

Contexto	Tipo de voz	Taxa de variação
UEn no início ou no meio da frase	TTS	62,15744
UEn no início ou no meio da frase	HUM	93,92609
Final das frases do GPS	TTS	-21,77255
Final das frases do GPS	HUM	-24,84031

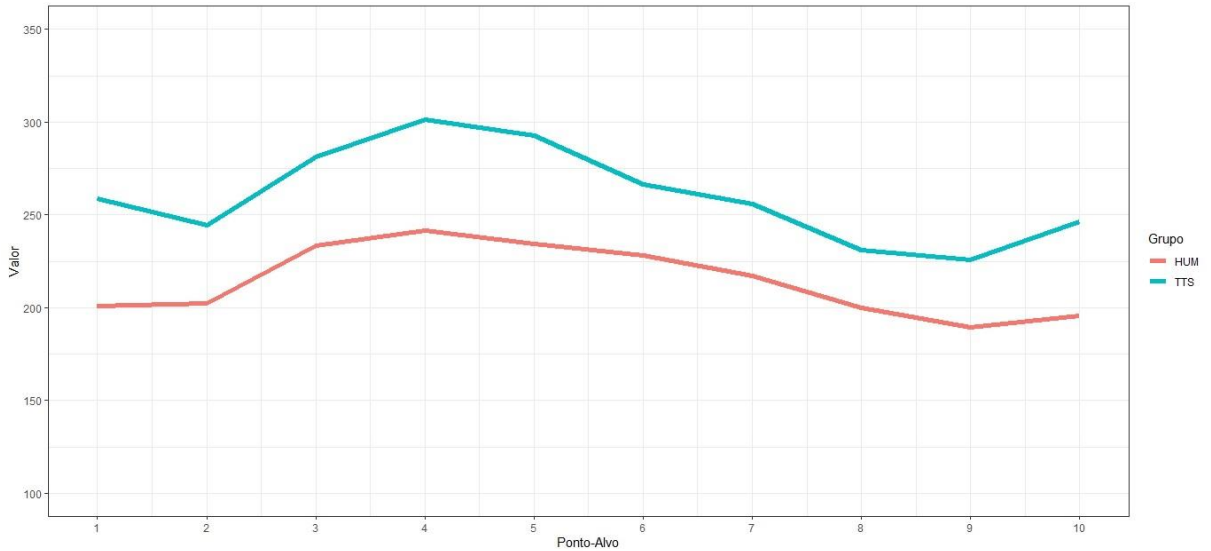
Fonte: elaborado pelo autor (2023)

Os testes t mostraram que não houve uma diferença significativa entre as médias dos dois grupos nas UEn do início e do meio da frase (p = 0.4877). Além disso, também não houve diferença significativa entre as médias de TTS e de HUM nos finais das frases (p = 0.9299).

4.4 Representação gráfica das UEn do início/meio das frases

A partir de 10 pontos-alvo de f0 obtidos pelo ProsodyPro, representamos graficamente as curvas das UEn que apareciam no início/meio das frases com as médias de TTS e de HUM. É possível observar semelhanças nas curvas dos dois grupos e certas diferenças entre os pontos 1 e 2 e entre os pontos 5, 6 e 7 (Gráfico 9).

Gráfico 9 - Representação gráfica das curvas da UEn no início ou no meio da frase

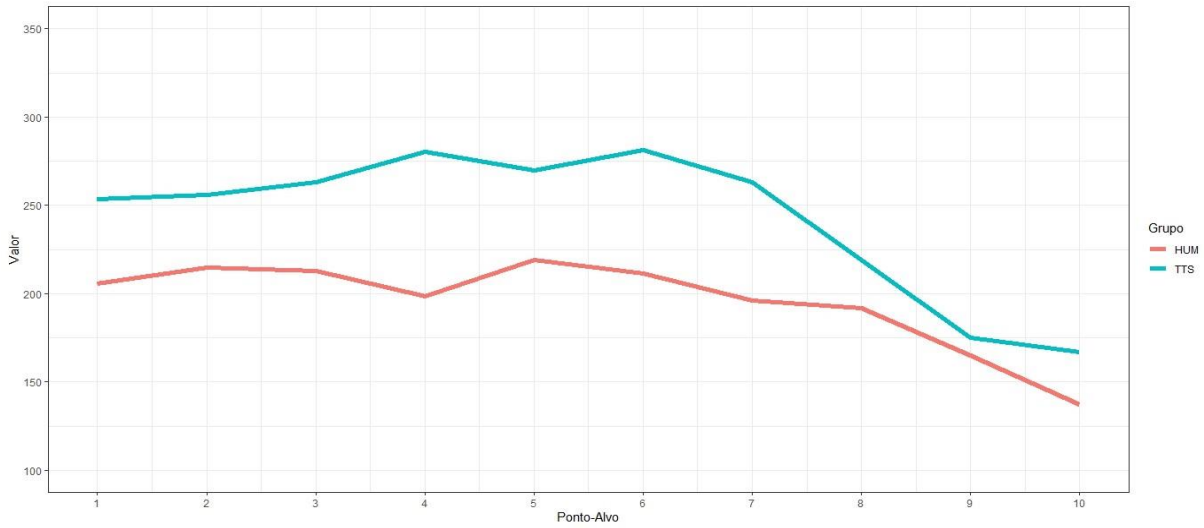


Fonte: elaborado pelo autor (2023)

4.5 Representação gráfica das UEn no final das frases

Também representamos graficamente as curvas das UEn que aparecem no final das frases com as médias de aplicativos e de humanos. As duas curvas apresentam semelhanças se observados dois picos evidentes e uma descida gradual na curva. No entanto, há diferenças: o primeiro e o segundo pico entoacional nos aplicativos é mais tardio do que os picos nos aplicativos. Além disso, há uma leve inclinação no fim da curva no grupo TTS, enquanto há uma descida total no grupo HUM (Gráfico 10).

Gráfico 10 - Representação gráfica das curvas da UEn no final da frase



Fonte: elaborado pelo autor (2023)

5 DISCUSSÃO

Nesta seção, apresentamos uma discussão dos resultados desta pesquisa, tendo em vista os objetivos descritos na introdução. Inicialmente, discutimos a análise acústica da f0, seguida pela análise das curvas de f0. Por fim, realizamos uma discussão geral dos achados, retomando a nossa hipótese geral de que os padrões entoacionais dos aplicativos são diferentes dos padrões dos humanos.

5.1 Discussão da análise acústica de f0

Um dos objetivos da presente pesquisa foi investigar as características acústicas da entoação de frases diretivas realizadas por síntese de fala em sistemas de GPS. Além disso, o estudo procurou comparar os resultados obtidos a partir da análise dessas frases sintetizadas com os resultados das análises de dados produzidos por seres humanos. Para isso, consideramos os valores de f0 – mínima, média, máxima – nas unidades entoacionais e no primeiro pico entoacional nos dados experimentais. A seguir, discutiremos os resultados dessas análises.

5.1.1 Discussão da análise acústica das unidades entoacionais

Ao investigar as UE, anotadas na camada UE+, é possível perceber que, dependendo da medida de f0 verificada, há diferenças ou semelhanças entre as quatro locutoras deste estudo. (Tabela 9).

Tabela 9 - Tabela-resumo de resultados entre as locutoras na camada UE+

Medida de f0	Locutoras com diferença significativa	Locutoras sem diferença significativa
F0_min	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ), (P2 e WZ).	(GM e WZ).
F0_mean	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ), (P1 e P2), (GM e WZ).	–
F0_max	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ).	(GM e WZ), (P1 e P2).

Fonte: elaborado pelo autor (2023)

Os resultados indicam que, independentemente da medida de f0 analisada, as médias do *Google Maps* e do *Waze* são diferentes de qualquer uma das duas participantes. Além disso, é possível verificar que a participante 1 e a participante 2 podem diferir entre as médias de f0_min e f0_mean, mas isso não acontece com a f0 máxima, em que os valores se aproximam, o que ocorre também entre os dois aplicativos.

Em relação à variável voz na camada UE+, é possível concluir que, ao compararmos as médias de f0_min de aplicativos e humanos, haverá diferenças nas UEn e nas UE1, mas haverá semelhanças nas UEa e na UE2. Uma possível explicação para esse achado é que as UEn contêm o verbo de ação, e a UE1 está próxima à UEn. Assim, é interessante notar que apenas as UEs periféricas (ou numa ponta ou na outra da frase) não apresentaram diferença significativa nas médias de f0_min, mas, em todas as medidas, a UEn é diferente entre HUM e TTS (Figura 6).

Figura 6 – Resultado da variável voz na camada UE+

Fonte: Elaborado pelo autor.

Outra possível interpretação para esse achado é que, ao dividir as quatro locutoras em dois grupos, as médias de f0_min podem chegar a valores mais próximas entre os dois grupos nas UEa e UE2 e valores mais distantes nas UEn e UE1. É importante notar que as UE possuem diferentes extensões de tamanho, o que pode levar a uma diferença também nas medidas. Ao analisar as médias de f0_mean e f0_max, pode-se concluir que as frases inteiras são acusticamente diferentes entre HUM e TTS, pois as quatro unidades apresentaram diferenças entre os dois grupos nessas duas medidas.

Além disso, os resultados sobre variação de f0 (= *f0_max – f0_min*) mostraram que as médias de aplicativos são significativamente diferentes dos humanos. A literatura indica que a variação de f0 está associada a emoções, como raiva, alegria e tristeza (VASSOLER & MEDEIROS, 2013; BHARADWAJ & ACHARJEE, 2023). Uma vez que a expressão de sentimentos na síntese de fala é uma preocupação atual (ROBINSON *et al*., 2019; ZHAO *et al*., 2020), o fato de a variação de f0 apresentar diferenças em nossos resultados evidencia um desafio para a síntese de fala. Esse aprimoramento pode ser alcançado por meio de modelos que geram f0 automaticamente através de informações obtidas em bancos de fala natural (MARQUES; UEDA; COSTA, 2022).

Além da frequência fundamental, outros aspectos prosódicos, como intensidade e duração, também são passíveis de análise. Seria importante investigar, por exemplo, se as frases proferidas pela síntese de fala possuem valores de intensidade particulares e se diferentes aspectos de duração, como taxa de elocução, duração da última sílaba tônica e tempo e duração de excursão de f0 diferem ou se assemelham à produção típica da fala produzida por humanos.

5.1.2 Discussão da análise acústica do primeiro pico das unidades entoacionais

Ao analisar o primeiro pico entoacional das unidades, é possível concluir que há proximidade entre as duas locutoras humanas, mas há diferenças entre qualquer uma das outras locutoras, independentemente da medida de f0 considerada (Tabela 10).

Tabela 10 - Tabela-resumo de resultados entre as locutoras na camada pUE+

Medida de f0	Locutoras com diferença significativa	Locutoras sem diferença significativa
F0_min	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ), (GM e WZ).	(P1 e P2).
F0_mean	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ), (GM e WZ).	(P1 e P2).
F0_max	(P1 e GM), (P2 e GM), (P1 e WZ), (P2 e WZ), (GM e WZ).	(P1 e P2).

Fonte: elaborado pelo autor (2023)

Esse resultado sobre o primeiro pico entoacional difere do resultado sobre as médias da camada discutida anteriormente. Ao analisar apenas o pico, é possível observar que, além das semelhanças entre as duas participantes, os dois aplicativos não apresentam semelhanças acústicas entre si em nenhuma das medidas de f0.

Em relação à variável voz, pode-se concluir que o primeiro pico das unidades entoacionais é diferente entre humanos e aplicativos nas UEa, UEn e UE1 nas três medidas de f0. A única unidade em que os valores se aproximam é a UE2 nas três medidas. Essa unidade já apresentou semelhança entre os dois grupos nos resultados da camada anterior, mas apenas na medida de f0_min. Isso sugere que o primeiro pico das UE2 possui um valor baixo se comparado com os das outras UE.

O primeiro pico entoacional, segundo Rocha (2016), pode caracterizar as atitudes do falante. Para o autor, os valores desse pico indicam se as frases ditas como ordem, por exemplo, podem ter uma atitude de referência, cortesia ou irritação. Essas atitudes podem interferir no conteúdo do enunciado e estabelecer relações interpessoais, como realizar uma fala mais polida ou arrogante, por exemplo (MORAES, *et al*., 2010).

Uma vez que nossos resultados sobre o primeiro pico mostraram distinções em três das quatro UEs analisadas, isso sugere que pode haver ou não uma mudança de atitude entre aplicativos e humanos na fala instrutiva. Descrever atitudes na fala diretiva foge do escopo do presente trabalho. Estudos futuros podem investigar se as vozes analisadas aqui apresentam, do ponto de vista da produção, uma atitude mais ríspida, próxima à da ordem, ou outro tipo de atitude. Isso pode ser estudado a partir da literatura sobre o papel da prosódia nas atitudes da fala diretiva (QUEIROZ, 2011; ROCHA, 2016).

5.2 Discussão da análise das curvas de f0

Além de analisar acusticamente a f0, esta pesquisa investigou as curvas das frases do GPS, com destaque para a UEn. Para isso, calculamos a taxa de variação das curvas e verificamos a excursão dos movimentos, a partir da diferença entre média inicial (mi) e ataque (a) no primeiro contorno, média final (mf) e (mi) no segundo contorno e último pico entoacional e penúltimo pico entoacional no terceiro contorno. Além disso, representamos as médias de 10 pontos-alvo de f0 das UEn que apareciam no início ou no meio da frase e das UEn que eram proferidas no final da frase.

5.2.1 Discussão da taxa de variação e da excursão dos movimentos

A análise do primeiro movimento entoacional mostrou que humanos e aplicativos realizam uma configuração ascendente de f0. Os resultados em hertz mostraram que há diferenças entre as duas vozes nesse primeiro movimento. Os resultados em semitons não indicaram diferença significativa.

No segundo movimento, ambas as falas produzem uma configuração descendente, mas os valores em hertz não se aproximam, enquanto as médias em semitons não sugerem uma diferença significativa entre uma voz sintética e uma voz humana. Nas medidas em hertz, a média das diferenças entre (mf) e (mi) de TTS (-44,46296) e a média dos HUM (-6,433962) já demonstram uma grande diferença de realização.

Em relação ao terceiro movimento, analisado a partir das UEn que aparecem no início ou no meio da frase, os resultados mostraram semelhanças na configuração ascendente de f0 tanto em humanos quanto nos aplicativos. O mesmo aconteceu no movimento final das frases, em que a configuração foi descendente em ambas as falas, mas houve diferença significativa na medida em hertz.

Os valores em semitons se valem de um cálculo logarítmico (BARBOSA, 2019), o que permite concluir que haverá ou não semelhanças entre aplicativos e humanos em relação à produção e à percepção da fala. Assim, o primeiro movimento entoacional pode gerar uma percepção de *pitch* mais aguda nos aplicativos, enquanto a produção do segundo movimento é que aparenta diferenças.

Uma possibilidade mais adequada para estudar a percepção dessas frases é testar a sua aceitabilidade. Experimentos de aceitabilidade podem revelar se usuários preferem uma voz mais empática e podem fornecer um *feedback* dos usuários sobre as melhorias dos sistemas para os desenvolvedores (WAGNER *et al*., 2019; JAMES *et al*., 2020). Isso é necessário, pois a aceitabilidade é um fator essencial para a integração de novas tecnologias, principalmente quando os usuários são idosos ou pessoas com pouca instrução (PORTET, *et al*., 2013). Esse teste também poderia envolver as condições de felicidade (ou sinceridade) tratadas na Teoria dos Atos de Fala (AUSTIN, 1990; SEARLE, 1991), uma vez que, se a voz do GPS é aceita, é provável que a síntese de fala alcance o objetivo de que o falante queira que o ato se realize.

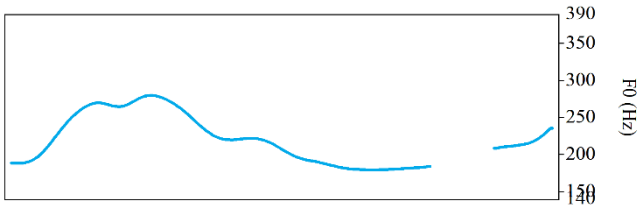
5.2.2 Discussão da representação gráfica das curvas

A representação gráfica das UEn que estão no início ou no meio das frases demonstra que há pontos em que as médias de f0 da fala sintética se aproximam da fala humana. As semelhanças podem ser percebidas quando se verifica um movimento de subida, uma descida gradual da curva e uma leve subida final, em que a curva se prepara para subir novamente.

No entanto, há diferenças entre o ponto-alvo 1 e 2 nessas UEn, pois os aplicativos variam os valores de f0, enquanto o movimento da fala dos humanos não apresenta uma variação considerável. Uma segunda diferença reside no fato de que, entre os pontos 5, 6 e 7, a curva correspondente à dos humanos realiza um movimento contínuo, enquanto os aplicativos apresentam valores diferentes, fazendo a descida da curva ficar instável.

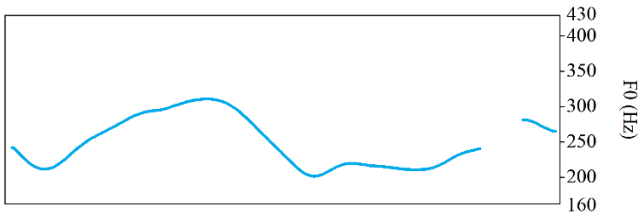
As UEn que estão no início ou no meio das frases representam a maior parte das UEn nas frases dos aplicativos de GPS. Criamos curvas reais[4] de "Vire à direita" na frase 11[5], realizadas pela participante 1 (Figura 7) e pelo Google Maps (Figura 8) para discutir alguns pontos importantes nessas curvas.

Figura 7 - UEn no início da frase realizada pela participante 01



Fonte: Elaborado pelo autor.

Figura 8 - UEn no início da frase realizada pelo Google Maps



Fonte: Elaborado pelo autor.

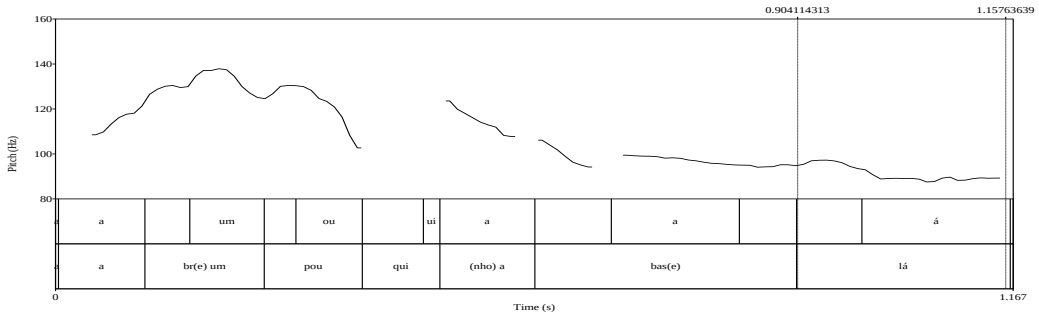
As duas curvas mostram que, após os primeiros picos, há uma queda de f0. No fim da curva, as locutoras realizam uma subida. Essas observações sugerem duas conclusões: (i) as

[4] As curvas de f0 foram extraídas a partir de um *script* do Praat disponível em: https://github.com/wendyelviragarcia/create_pictures.
[5] Frase 11 do Google Maps: "Vire à direita na Rua Doutor José Bernardino Carvalho Leite".

curvas reafirmam os achados de Almeida (2017) e de Kent & Read (2015), de que essa queda marca uma unidade entoacional no meio da frase, pois a subida indica que o falante vai continuar a sentença; (ii) os aplicativos seguem esse padrão final da fala natural.

Não obstante, no início das curvas, existem diferenças. Os aplicativos parecem realizar o que é considerado na literatura como "vale", em razão de seu formato, antes do primeiro pico, enquanto a participante 1 realiza um início de frase de forma nivelada, antes do pico, tal qual os contornos entoacionais da instrução no português brasileiro (Figura 9). Isso sugere que os aplicativos não seguem esse padrão inicial da fala natural.

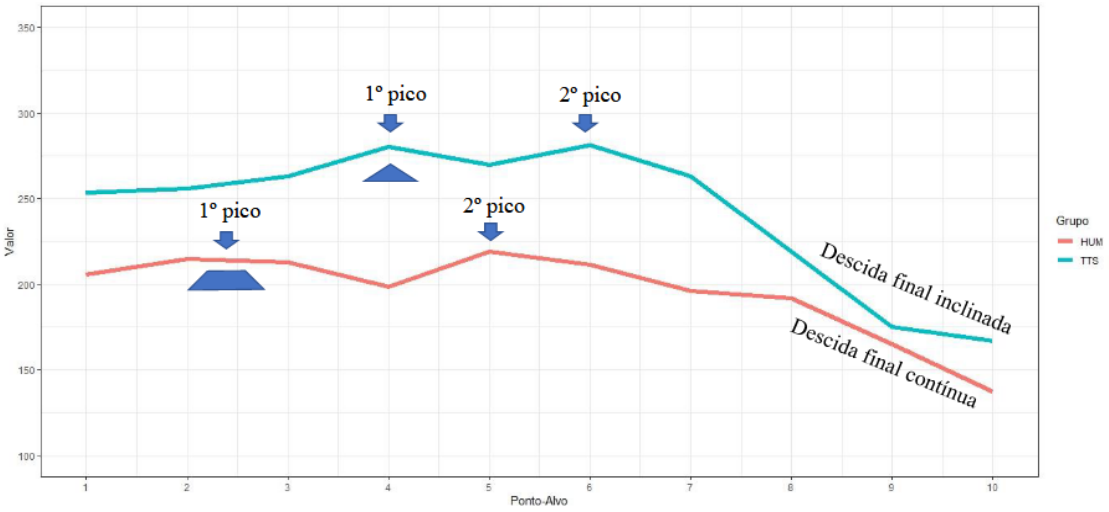
Figura 9 - Exemplo típico de contorno entoacional da instrução no português brasileiro



Fonte: Rocha (2013).

A representação gráfica das UEn no final da frase revela semelhanças entre as duas curvas, pois há dois picos iniciais e uma descida final que deixa evidente a configuração descendente indicada pela excursão de movimento. Porém, as curvas das UEn finais dos humanos são diferentes das curvas dos aplicativos (Gráfico 11), uma vez que o primeiro pico apresenta um efeito platô na fala das participantes, ou seja, há uma subida de f0, uma estabilidade nas medidas e depois uma descida. Além disso, os valores de f0 são maiores nos aplicativos, e a descida final é contínua nos humanos, enquanto possui uma leve inclinação na fala sintética.

Gráfico 11 - Diferenças entre TTS e HUM nos picos e no movimento final nas UEn finais



Fonte: elaborado pelo autor (2023)

Tanto o formato dos picos quanto a descida final no enunciado podem ter um impacto na percepção dos usuários do aplicativo, pois são diferentes na curva da fala natural. Tal resultado indica novamente a necessidade de estudos futuros realizarem um teste perceptual sobre a fala sintética do GPS. A percepção de picos em enunciados falados humanos afeta a compreensão da fala, uma vez que um longo platô pode fazer um pico soar mais alto, além de os ouvintes atribuírem mais atenção a picos no processamento da fala, por exemplo (JEON & HEINRICH, 2022).

O primeiro e o segundo pico nas UEn finais dos aplicativos estão localizados mais tarde do que os picos entoacionais dos humanos. No entanto, essa diferença entre os grupos não acontece nas UEn que estão no início ou no meio da frase. Uma possível explicação para isso pode estar relacionada com o acento ou o ataque. Ao investigar o modelo de entoação usado no Sistema de Multilíngua da *Bell Labs*, Santen *et al*. (1998) evidenciaram que o pico entoacional acontece mais tarde em sílabas acentuadas e mais cedo em grupos de acento monossilábicos.

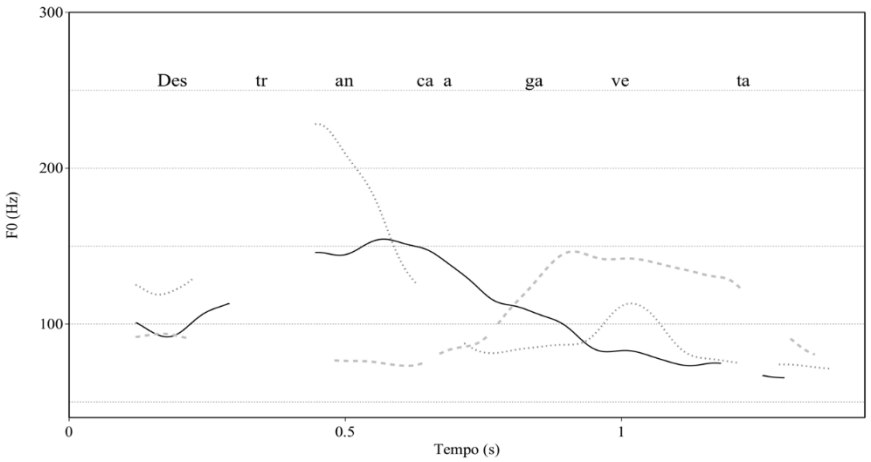
No sueco, diferentes variantes dessa língua podem mostrar um alinhamento temporal do pico mais tardio a depender do acento, indicando que essas análises prosódicas são valiosas para a síntese de fala (SCHÖTZ *et al*., 2010). Pesquisas futuras podem observar se há relação entre o acento e a localização temporal do pico na síntese de fala e nas regiões do português brasileiro, sobretudo na fala instrutiva.

A representação gráfica das curvas e a análise acústica revelam a presença de valores maiores de f0 nos aplicativos. Os valores médios de f0 para vozes tipicamente do sexo feminino variam entre 150 a 250 Hz (GUIMARÃES, *et al*., 2003). Re *et al*. (2012) afirmam que homens

preferem mais vozes femininas agudas, enquanto mulheres preferem vozes masculinas graves, o que demonstra uma análise da percepção de *pitch* e de atratividade vocais. Se as vozes femininas são preferencialmente mais aceitas quando são mais agudas, o fato de o GPS ter valores mais altos de f0 do que as participantes, em nossa pesquisa, revela que isso talvez seja intencional.

A pouca aproximação nas curvas evidenciada nos dois primeiros picos e na descida final das UEn finais sugere que a fala dos aplicativos e dos humanos, visualmente, possuem diferenças entoacionais. Apesar dessas diferenças, as duas curvas se distanciam de contornos típicos de outros atos diretivos, como ordem, pedido e sugestão (Figura 10).

Figura 10 - Contorno melódico dito como ordem (linha preta contínua), pedido (linha cinza pontilhada) e sugestão (linha cinza tracejada)



Fonte: Moraes; Rilliard (2022)

5.3. Discussão geral dos resultados

Esta pesquisa buscou analisar e comparar dados de fala de aplicativos e de humanos, a partir de uma análise acústica de três medidas de f0, da variação de f0 e do primeiro pico entoacional. Adicionalmente, este estudo examinou a configuração de f0 e comparou as curvas entoacionais, utilizando uma representação gráfica.

As frases analisadas nesta pesquisa foram realizadas por locutoras dos aplicativos *Google Maps* e *Waze* e por duas locutoras do português brasileiro na variedade alagoana. A análise foi baseada na frequência fundamental, que desempenha um papel importante na síntese da fala e influencia diretamente a naturalidade (JANYOI; SERESANGTAKUL, 2020b). O objetivo dessas frases é induzir o ouvinte a tomar uma determinada decisão, de acordo com a Teoria dos Atos de Fala, cuja realização desses atos depende da entoação, que é responsável

por gerar diferentes atos de fala diretivos, como ordem, instrução, pedido e sugestão. Assim, era necessário verificar nesta pesquisa se a entoação do GPS se aproximava da fala natural.

Os nossos resultados indicaram diferenças na entoação entre aplicativos e humanos, uma vez que houve diferença significativa na variação de f0. Além disso, concluiu-se que as frases constituídas pelas unidades entoacionais são acusticamente diferentes entre os dois grupos, pois as quatro unidades analisadas apresentaram diferenças nas médias de f0 média e f0 máxima. A análise das curvas mostrou que a configuração de f0 dos aplicativos é semelhante à dos humanos, mas há diferenças nas medidas em hertz e semitons e na representação gráfica das curvas.

Nossos achados apresentam algumas diferenças significativas entre aplicativos e humanos, mas nossa pesquisa possui limitações, como a ausência de um inventário tonal e a falta de um teste de percepção da fala sintética. Trabalhos futuros podem investigar se usuários do GPS sentem que os processos de compreensão são prejudicados ou se sentem desmotivados a ouvir a voz dos aplicativos por um longo caminho, por exemplo. Ao investigar a percepção da fala sintética de sistemas TTS, Greene, Logan e Pisoni (2012) concluem que é necessário pesquisar como humanos lidarão ao ouvir longas passagens de fala sintética, se haverá tédio ou perda da concentração, ou se a compreensão será afetada pela qualidade dessa voz.

Esses estudos sobre percepção são importantes, pois indicarão se as pessoas já estão acostumadas com a fala sintética e se a compreensão e a aceitação dependem de outros fatores, como nível de instrução do ouvinte ou idade, por exemplo. Para Noah *et al*. (2021), falantes nativos preferem vozes humanas, mas o fato de saberem anteriormente que ouvirão uma voz computadorizada pode permitir que eles compreendam que essa fala apresenta limitações.

Esta dissertação destaca algumas diferenças entre aplicativos e humanos e ressalta a importância desses estudos de percepção, já que uma fala sintética mais natural e aceitável é particularmente relevante em situações que envolvem tomadas de decisão, como no caso do uso do GPS. Enunciados que possam gerar dificuldades de aceitação ou de compreensão desses aplicativos podem resultar em experiências ruins para o usuário ou mesmo consequências nocivas, seja no trânsito ou em lugares desconhecidos pelo usuário.

Além das diferenças, o fato de também haver semelhanças entre as falas analisadas nesta dissertação já é um indício de que a voz produzida pelos aplicativos apresenta avanços em sua realização. Esses avanços recentes ajudaram a fala sintética a ir em direção a uma fala mais natural, além de uma simples imitação robótica da fala humana (NOAH *et al*., 2021). Essa aproximação é verificada principalmente na configuração de f0 em nossos achados.

Para a literatura, a instrução no português revela uma configuração ascendente, descendente e descendente de f0 (ROCHA, 2013). Essa configuração aparece em nossos resultados tanto no grupo dos humanos quanto no grupo dos aplicativos se considerarmos o movimento final das UEn que aparecem no final das frases. No entanto, como a maioria das falas diretivas que têm o verbo de ação aparecem no início ou no meio das sentenças, a configuração mais comum das UEn do GPS, nos dois grupos, é ascendente, descendente e ascendente de f0.

Partimos da hipótese inicial da pesquisa de que os padrões entoacionais do GPS são diferentes das vozes humanas. Nossa hipótese foi parcialmente confirmada, uma vez que algumas das semelhanças discutidas acima não devem ser ignoradas. Ao mesmo tempo, pode-se concluir que as frases inteiras do GPS são acusticamente diferentes de frases realizadas por humanos se considerarmos que houve diferença significativa na variação de f0 (= *f0_max – f0_min*) e na representação gráfica das curvas.

As diferenças entre os contornos entoacionais podem ser diminuídas, a partir da modelagem automática de contornos de f0, como os Modelos Ocultos de Markov (DALL *et al.*, 2016; GHONE *et al*., 2017; JANYOI & SERESANGTAKUL, 2020b). É importante que essas técnicas levem em conta a relação existente entre a entoação e a fala diretiva, evidenciada nesta pesquisa e também em trabalhos anteriores na Linguística. No entanto, os sistemas TTS ainda encontram desafios, pois, para que sejam treinados, é preciso haver um conjunto de dados grande e de qualidade, um recurso nem sempre disponível para todas as línguas (AHMAD *et al*., 2021; CHOI *et al*., 2022).

6 CONSIDERAÇÕES FINAIS

Esta dissertação buscou analisar, descrever e comparar a entoação de vozes de aplicativos de GPS e de duas falantes nativas do português na variedade alagoana do sexo feminino. Buscamos a literatura a fim de compreender a relação que a entoação estabelece entre a síntese de fala e a fala diretiva, descrevemos os procedimentos metodológicos para analisar os dados e apresentamos e discutimos os resultados.

Os achados dessa dissertação mostraram que há diferenças acústicas significativas entre humanos e aplicativos na variação de f0 (= *f0_max – f0_min*) e nas médias de f0 média e f0 máxima nas quatro unidades entoacionais analisadas. Além disso, foram observadas diferenças acústicas entre os dois grupos no primeiro pico entoacional em três das quatro unidades entoacionais das frases e diferenças na representação gráfica das curvas. No entanto, os resultados mostraram semelhanças entre aplicativos e humanos, na configuração de f0 (ascendente, descendente, descendente). Assim, as análises acústicas e de contorno de f0 confirmam parcialmente a hipótese deste trabalho, uma vez que há diferenças e semelhanças entre as duas falas.

Considerando que a Teoria dos Atos de Fala sugere que a entoação é um aspecto importante para a comunicação, é necessário continuar aprimorando a fala dos aplicativos e verificar níveis de aceitabilidade, naturalidade e de inteligibilidade, que podem afetar o significado e a interpretação da informação fornecida. Entre muitos critérios que são considerados atualmente para aprovação de qualquer sistema TTS, a naturalidade e a compreensão apresentam-se como os principais pontos de avaliação (CHEMNAD & OTHMAN, 2023). Uma vez que a presente pesquisa revelou uma diferença significativa na variação de f0 entre aplicativos e humanos, é necessário desenvolver novos métodos e técnicas que ajudem a criar uma síntese ideal para o português, aprimorando a naturalidade dessa fala.

A pesquisa atingiu seus objetivos, ao analisar a acústica da síntese de fala do GPS, descrever os contornos dessa voz e compará-la com a entoação dos humanos. Esta dissertação fornece uma contribuição para a entoação, uma vez que fornece pistas para a descrição de uma variedade pouco estudada: o português brasileiro na variedade alagoana. Uma vez que a presente pesquisa evidenciou lacunas, é importante continuar o trabalho acerca da síntese de fala, principalmente a partir das contribuições da Linguística.

A análise do contorno entoacional deste trabalho tem como foco as unidades entoacionais nucleares. Pensamos em ampliar este estudo futuramente, a fim de descrever todas

as unidades entoacionais das frases do GPS, a partir de um inventário tonal, utilizando os *scripts* Momel/Intsint (HIRST, 2007) e ProsodyPro (XU, 2013). Esse experimento forneceria uma visão geral das frases dos aplicativos, uma vez que elas possuem diversas estruturas que podem apresentar distinções prosódicas mais precisas entre a fala de humanos e de máquinas.

A presente pesquisa fornece informações sobre padrões acústicos e descritivos da f0 de máquinas e de humanos e contribui para o entendimento da tecnologia da fala, para a área da entoação e para o estudo das frases diretivas. Pesquisas futuras que envolvam profissionais da linguística e da computação podem ampliar sua importância, através de uma descrição entoacional de falas diretivas de outras regiões do país além do português na variedade alagoana.

O avanço promissor da inteligência artificial trará sempre novos desafios para a prosódia, uma vez que a síntese de fala terá que lidar com diferentes realizações de contornos entoacionais presentes na fala humana, que possui um caráter irrestrito. Além disso, questões éticas e forenses também deverão ser enfrentadas.

REFERÊNCIAS

AGUILAR, L. “La entonación.” In: ALCOBA, S. **La expresión oral**. Barcelona: Ariel, 2000, p. 115-141.

AHMAD, A.; SELIM, M. R.; IQBAL, M. Z.; RAHMAN, M. S. SUST TTS Corpus: A phonetically-balanced corpus for Bangla text-to-speech synthesis. **Acoustical Science and Technology**, v. 42, n. 6, p. 326-332, 01 nov. 2021. Disponível em: https://doi.org/10.1250/ast.42.326. Acesso em: 11 maio 2023.

ALMEIDA, A. N. S. **Análise prosódica de agrupamentos numéricos no português do Brasil**. Tese (Doutorado em Letras e Linguística). Universidade Federal de Alagoas, Maceió, 2017.

AUSTIN, J. L. **Quando dizer é fazer. Palavras e ação**. Porto Alegre: Artes Médicas, 1990.

AZUMA, S. O. **As estratégias de atos diretivos no ambiente corporativo na língua portuguesa falada na região de Curitiba e na língua japonesa falada por expatriados**. Dissertação (Mestrado em Linguística), Universidade Federal do Paraná, Curitiba, 2014.

BALLESTEROS, M.; WANNER, L. A Neural Network Architecture for Multilingual Punctuation Generation. **Association for Computational Linguistics**, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, Texas, p. 1048-1053, 2016. Disponível em: http://dx.doi.org/10.18653/v1/D16-1111. Acesso em: 8 fev. 2021.

BARBOSA, D. S. **Análise e proposição de modelos de síntese de fala para integração ao framework FIVE**. Dissertação (Mestrado em Engenharia de Computação), Universidade de Pernambuco, Recife, 2016.

BARBOSA, P. A. Revelar a estrutura rítmica de uma língua construindo máquinas falantes: pela integração de ciência e tecnologia de fala. *In*: SCARPA, E. M. (org.). **Estudos de prosódia**. Campinas, SP: Editora da Unicamp, 1999.

BARBOSA, P. A. **Prosódia**. São Paulo: Parábola, 2019.

BARRETO, R. S. C. **Análise entonacional de atos de fala diretivos na animação “Metegol”**. Monografia (Licenciatura em Letras Espanhol). Universidade Federal da Paraíba, João Pessoa, 2019.

BARTH-WEINGARTEN, D. From “intonation units” to cesuring – an alternative approach to the prosodic-phonetic structuring of talk-in-interaction. *In*: REED, B. S.; RAYMOND, G. (eds.). **Units of Talk – Units of Action**. Amsterdam: John Benjamins, 2013, p. 91-124.

BECKMAN, M.; HIRSCHBERG, J. **The ToBI annotation conventions**. Online MS, 1994. Disponível em: http://www.ling.ohio-state.edu/~tobi/ame_tobi/annotation_conventions.html. Acesso em: 03 maio 2023.

BECKMAN, M.; PIERREHUMBERT, J. Intonational structure in Japanese and English. **Phonology Yearbook**, v. 3, n. 1, p. 255-309, 1986.

BHARADWAJ, S.; ACHARJEE, P. B. Exploring human voice prosodic features and the interaction between the excitation signal and vocal tract for Assamese speech. **International Journal of Speech Technology**, v. 26, n. 1, p. 77-93, 2023. Disponível em: https://doi.org/10.1007/s10772-021-09946-5. Acesso em: 01 jun. 2023.

BILLMYER, K.; VARGHESE, M. Investigating instrument-based pragmatic variability: effects of enhancing discourse completion tests. **Applied Linguistics**, v. 21, n. 4, p. 517 – 552, 2000.

BIN, P. R.; MOTA, M. B. Pré-registro de estudos na Linguística Experimental. **Cadernos de Linguística**, v. 3, n. 1, e616, 2022. Disponível em: https://doi.org/10.25189/2675-4916.2022.v3.n1.id616. Acesso em: 30 jun. 2022.

BODOLAY, A. N. **Pragmática da entonação: a relação prosódia/contexto em atos diretivos no Português**. (Tese de Doutorado). Universidade Federal de Minas Gerais, Belo Horizonte, 2009. Disponível em: https://repositorio.ufmg.br/handle/1843/ALDR-7R5NRD Acesso em: 30 out. 2020.

BODOLAY, A. N. O papel da prosódia em enunciados de ordens e pedidos. **Revista Científica Vozes dos Vales**. UFVJM – MG – Brasil, n. 6 – Ano III, 2014. Disponível em: http://site.ufvjm.edu.br/revistamultidisciplinar/files/2014/10/O-papel-da-pros%C3%B3dia-em-enunciados-de-ordens-e-pedidos.pdf Acesso em: 30 ago. 2021.

BODOLAY, A. N. Prática de ensino de português como língua estrangeira. **Revista (Com) Textos Linguísticos**. Vitória, v. 14, n. 29, 2020. Disponível em: https://doi.org/10.47456/cl.v14i29.32139 Acesso em: 30 ago. 2021.

BOERSMA, P; WEENINK, D. **Praat**, 1993-2023. Disponível em: http://www.fon.hum.uva.nl/praat/. Acesso em: 23 mar. 2021.

BOLINGER, D. L. **Intonation and its Uses**: Melody in Grammar and Discourse. Stanford: Stanford University Press, 1989.

BRAGA, D. Máquinas falantes: Novos paradigmas da língua e da linguística. **Colóquio Política Linguística**, 2007. Disponível em: http://download.microsoft.com/download/A/0/B/A0B1A66A-5EBF-4CF3-9453-4B13BB027F1F/ColoquioPoliticaLinguistica_2007.pdf. Acesso em: 07 ago. 2021.

BRAGA, G.; FROTA, S.; SVARTMAN, F. R. F. Português guineense e português europeu: um estudo preliminar sobre a percepção das suas diferenças entoacionais. **Revista do GEL**, v. 18, n. 3, 2021. Disponível em: http://dx.doi.org/10.21165/gel.v18i3.3164. Acesso em: 26 ago. 2022.

CARVALHO, T. G. **Uso de vírgulas em textos do Ensino Fundamental II**: um estudo longitudinal. Dissertação (Mestrado em Estudos Linguísticos). Universidade Estadual Paulista Júlio de Mesquita Filho, São José do Rio Preto, 2019.

CASANOVA, E. **Síntese de voz aplicada ao português brasileiro usando aprendizado profundo.** Monografia (Bacharelado em Computação). Universidade Tecnológica Federal do Paraná, Departamento Acadêmico de Computação, Medianeira, 2019.

CAVALCANTE, D. L. **Uma análise comparativa dos codificadores/decodificadores de voz para comunicações digitais**. Dissertação (Mestrado em Engenharia Elétrica). Universidade Federal de Pernambuco, Centro de Tecnologia e Geociências, Recife, 2009.

CHAFE, W. Pontuação e a Prosódia da Linguagem Escrita. **Written Communication**, v. 5, n. 4, p. 395-426, 1988. Disponível em: https://doi.org/10.1177/0741088388005004001. Acesso em: 13 mar. 2023.

CHBANE, D. T. **Desenvolvimento de sistema para conversão de textos em fonemas no idioma português.** Dissertação (Mestrado em Engenharia). Universidade de São Paulo, Escola Politécnica, São Paulo, 1994.

CHEMNAD, K.; OTHMAN, A. Advancements in Arabic Text-to-Speech Systems: A 22-Year Literature Review. **IEEE Access**, v. 11, p. 30929-30954, 2023. Disponível em: https://doi.org/10.1109/ACCESS.2023.3260844. Acesso em: 12 jun. 2023.

CHEN, S.; HWANG, S. WANG, Y. An RNN-based prosodic information synthesizer for Mandarin text-to-speech. **IEEE Transactions on Speech and Audio Processing**, v. 6, n. 3, p. 226-239, May 1998, Disponível em: https://doi.org/10.1109/89.668817. Acesso em: 08 ago. 2021.

CHEN, J.; YANG, H.; WU, X.; MOORE, B. C.J. The effect of F0 contour on the intelligibility of speech in the presence of interfering sounds for Mandarin Chinese. **The Journal of the Acoustical Society of America**, v. 143, n. 2, p. 864-877, 2018. Disponível em: https://doi.org/10.1121/1.5023218. Acesso em: 22 maio 2021.

CHOI, B. J.; JEONG, M.; LEE, J. Y.; KIM, N. S. SNAC: Speaker-Normalized Affine Coupling Layer in Flow-Based Architecture for Zero-Shot Multi-Speaker Text-to-Speech. **IEEE Signal Processing Letters**, v. 29, p. 2502-2506, 2022. Disponível em: https://doi.org/10.1109/LSP.2022.3226655. Acesso em: 11 maio 2023.

COOPER, E. L. **Text-to-Speech Synthesis Using Found Data for Low-Resource Languages**. Thesis (Doctor of Philosophy in the Graduate School of Arts and Sciences), Columbia University, 2019.

CORREA, C. P. V. **Descrição e análise fonética do dialeto de Belém do Pará orientadas à incorporação de conhecimento fonético à conversão de texto em fala.** Dissertação (Mestrado em Processamento de Linguagem Natural e Indústrias das Línguas). Universidade do Algarve, Universitat Autònoma de Barcelona, 2010.

COUPER-KUHLEN, E. **An Introduction to English Prosody**. Edward Arnold, London, 1986.

CUMBERS, B. A. **Perceptual correlates of acoustic measures of vocal variability**. Thesis (Master of Science). University of Wisconsin-Milwaukee, 2013

DALL, R.; HASHIMOTO, K.; OURA, K.; NANKAKU, Y.; TOKUDA, K. Redefining the Linguistic Context Feature Set for HMM and DNN TTS Through Position and Parsing. **INTERSPEECH 2016**, Sep. 8-12, San Francisco, USA, 2016. Disponível em: http://dx.doi.org/10.21437/Interspeech.2016-399. Acesso em: 06 maio 2021.

DUSTERHOFF, K.; BLACK, A. W. Generation F0 contours for speech synthesis using the tilt intonation theory. **Proc. ESCA Workshop on Intonation**, p. 107-110, Athens, Greece. 1997. Disponível em: https://era.ed.ac.uk/handle/1842/1228. Acesso em: 08 ago. 2021.

DUTOIT, T. **An introduction to text-to-speech synthesis**. Kluwer Academic Publishers, 1997.

EGASHIRA, F. **Síntese de voz a partir de texto para a língua portuguesa**. Dissertação (Mestrado em Engenharia Elétrica). Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica, Campinas, 1992.

ESCANDELL-VIDAL, M. V. **Introducción a la pragmática**. Barcelona: Anthropos, Madrid, 1993.

FILHO, P. E. O. **Desenvolvimento de regras de pronúncia para a síntese de fala em Língua Portuguesa**. (Dissertação de Mestrado). Universidade Federal do Rio Grande do Sul, Porto Alegre, 2002. Disponível em: https://lume.ufrgs.br/handle/10183/1631 Acesso em: 27 out. 2020.

FROTA, S.; CRUZ, M.; FERNANDES-SVARTMAN, F.; COLLISCHONN, G.; FONSECA, A.; SERRA, C.; OLIVEIRA, P.; VIGÁRIO, M. Intonational variation in Portuguese: European and Brazilian varieties. *In*: FROTA, S.; PRIETO, P. (eds.), **Intonation in Romance**. Oxford: Oxford University Press, 2015, p. 235-283.

FROTA, S.; MORAES, J. Intonation of European and Brazilian Portuguese. *In*: WETZELS, W. L.; MENUZZI, S.; COSTA, J. (eds.), **The Handbook of Portuguese Linguistics**. New Jersey, EUA: John Wiley & Sons, Inc., 2016, p. 141-166.

FROTA, S.; VIGÁRIO, M. Aspectos de prosódia comparada: ritmo e entoação no PE e no PB. **Actas do XV Encontro da Associação Portuguesa de Linguística**, v. 1, p. 533-555, 2000. Disponível em: http://labfon.letras.ulisboa.pt/SonseMelodias/APLPEPB.pdf. Acesso em: 03 maio 2023.

FUJISAKI, H. Prosody, models, and spontaneous speech. IN: SAGISAKA, Y.; CAMPBELL, N.; HIGUCHI, N. (edits). **Computing Prosody: Computational Models for Processing Spontaneous Speech**. New York, Springer, 1997.

GHONE, A. S.; NERPAPAR, R.; KUMAR, P.; BABY, A.; SHANMUGAM, A.; SASIKUMAR, M.; MURTHY, H. A. TBT (Toolkit to Build TTS): A High Performance Framework to build Multiple Language HTS Voice. **INTERSPEECH 2017: Show & Tell Contribution**, Aug. 20–24, 2017, Stockholm, Sweden, 2017. Disponível em: https://www.isca-speech.org/archive/Interspeech_2017/pdfs/2042.PDF. Acesso em: 23 abr. 2021.

GOMES DA SILVA, C.; COUTO, L. R.; PINTO, M. S. Pedidos de Informação e Pedidos de Ação em Português do Brasil, fala carioca e em Espanhol Europeu, fala madrilena: variantes ou padrões entonacionais distintos? **Anais do Congresso Brasileiro de Prosódia**, v. 1, 2011. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/article/view/1268/1381 Acesso em: 30 ago. 2021.

GOMES DA SILVA, C.; MIRANDA, L. S.; CARNAVAL, M.; CUNHA, C. Z. A entoação da ordem no português do Brasil: uma descrição dialetal a partir do *Corpus* ALiB. **Journal of Speech Sciences**, v. 5, n. 2, p. 29-45, 2016. Disponível em: https://doi.org/10.20396/joss.v5i2.15063 Acesso em: 30 ago. 2021.

GOMES DA SILVA, C.; CARNAVAL, M.; MORAES, J. A. Atos de fala diretivos em português e em espanhol: uma análise acústica comparativa. **Entrepalavras**, Fortaleza, v. 10, n. 1, p. 326-345, jan-abr, 2020. Disponível em: http://dx.doi.org/10.22168/2237-6321-11751. Acesso em: 30 ago. 2021.

GOMES, L. C. T. **Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras.** Dissertação (Mestrado em Engenharia Elétrica). Unicamp, Faculdade de Engenharia Elétrica e de Computação, 1998.

GREENE, B. G.; LOGAN, J. S.; PISONI, D. B. Perception of synthetic speech produced automatically by rule: Intelligibility of eight text-to-speech systems. **Behavior Research Methods, Instruments, & Computers**, v. 18, n. 2, p. 100-107, 1986. Disponível em: https://doi.org/10.3758/BF03201008. Acesso em: 03 jun. 2023.

GUIMARÃES, I.; BARROS, E.; GAMA, I.; BEIRÃO, C. J. A Frequência Fundamental da voz de Adultos. Lisboa: **Revista Portuguesa de Otorrinolaringologia e Cirurgia Cérvico-Facial**, 2003. Disponível em: https://www.researchgate.net/publication/259892144_A_FREQUENCIA_FUNDAMENTAL_DA_VOZ_DE_ADULTOS. Acesso em: 03 jun. 2023.

HIROSE, K. Modeling of fundamental frequency contours for HMM-based speech synthesis: Representation of fundamental frequency contours for statistical speech synthesis. **IEEE 13th International Conference on Signal Processing** (ICSP), 2016, p. 171-176, Disponível em: https://doi.org/10.1109/ICSP.2016.7877818. Acesso em: 18 jul. 2021.

HIRST, D. A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. **Proceedings 16th International Congress of Phonetic Sciences (ICPhS, XVI)**, Saarbrücken, Germany, p. 1233-1236, 2007.

HIRST, D. **Analyse tier PRAAT script**, 2012.

HIRST, D.; DI CRISTO, A. **Intonational Systems, a survey of twenty languages.** Cambridge: Cambridge University Press, 1998.

INOUE, K.; HARA, S.; ABE, M.; HOJO, N.; IJIMA, Y. An investigation to transplant emotional expressions in DNN-based TTS synthesis. **Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)**, 2017, p. 1253-

1258, Disponível em: https://doi.org/10.1109/APSIPA.2017.8282231. Acesso em: 23 abr. 2021.

JAMES, J.; BALAMURALI, B. T.; WATSON, C. I.; MACDONALD, B. Empathetic speech synthesis and testing for healthcare robots. **International Journal of Social Robotics**, v. 13, p. 2119-2137, 2021. Disponível em: https://doi.org/10.1007/s12369-020-00691-4. Acesso em: 08 jun. 2023.

JANYOI, P.; SERESANGTAKUL, P. F0 modeling for Isarn speech synthesis using deep neural networks and syllable-level feature representation. **The International Arab Journal of Information Technology**, v. 17, n. 6, nov. 2020a. Disponível em: http://dx.doi.org/10.34028/iajit/17/6/9. Acesso em: 22 maio 2021.

JANYOI, P.; SERESANGTAKUL, P. Tonal contour generation for Isarn speech synthesis using deep learning and sampling-based F0 representation. **Applied Sciences**, v. 10, n. 18, 2020b. Disponível em: https://doi.org/10.3390/app10186381. Acesso em: 06 maio 2021.

JEON, H.; HEINRICH, A. Perceptual asymmetry between pitch peaks and valleys. **Speech Communication**, v. 140, p. 109-127, 2022. Disponível em: https://doi.org/10.1016/j.specom.2022.04.001. Acesso em: 01 jun. 2023.

KAMEOKA, H.; YOSHIZATO, K.; ISHIHARA, T.; KADOWAKI, K.; OHISHI, Y.; KASHINO, K. Generative Modeling of Voice Fundamental Frequency Contours. **IEEE/ACM Transactions on Audio, Speech, and Language Processing**, v. 23, n. 6, p. 1042-1053, June 2015. Disponível em: https://doi.org/10.1109/TASLP.2015.2418576. Acesso em: 18 jul. 2021.

KENT, R. D.; READ, C. **Análise acústica da fala**. Tradução Alexsandro Rodrigues Meireles. São Paulo: Cortez, 2015.

KLIMKOV, V.; MOINET, A.; NADOLSKI, A.; DRUGMAN, T. Parameter Generation Algorithms for Text-To-Speech Synthesis with Recurrent Neural Networks. **IEEE Spoken Language Technology Workshop** (SLT), 2018, p. 626-631, Disponível em: https://doi.org/10.1109/SLT.2018.8639626. Acesso em: 18 jul. 2021.

KLIMKOV, V.; NADOLSKI, A.; MOINET, A.; PUTRYCZ, B.; BARRA-CHICOTE, R.; MERRITT, T.; DRUGMAN, T. Phrase Break Prediction for Long-Form Reading TTS: Exploiting Text Structure Information. **Proc. Interspeech**, p. 1064-1068, 2017. Disponível em: http://dx.doi.org/10.21437/Interspeech.2017-419. Acesso em: 25 maio 2021.

KLIMKOV, V.; RONANKI, S.; ROHNKE, J.; DRUGMAN, T. Fine-grained robust prosody tranfer for single-speaker neural text-to-speech. **Interspeech**, 2019**.** Disponível em: https://arxiv.org/abs/1907.02479. Acesso em: 23 abr. 2021.

KORIYAMA, T.; KOBAYASHI, T. Prosody generation using frame-based Gaussian process regression and classification for statistical parametric speech synthesis. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2015, p. 4929-4933, DOI: https://doi.org/10.1109/ICASSP.2015.7178908. Acesso em: 18 jul. 2021.

LADD, D. R. Constraints on the gradient variability of pitch range (or) pitch level 4 lives! *In*: KEATING, P. (Ed.), **Papers in Laboratory Phonology III**. Cambridge, UK: Cambridge University Press, 1993, p. 43-63.

LADD, D. R. **Intonational Phonology**. 2nd ed. Cambridge: Cambridge University Press, 2008.

LANGARANI, M. S. E.; SANTEN, J. V. Foot-based intonation for text-to-speech synthesis using neural networks. **Conference: Speech Prosody**, May 2016. Disponível em: http://dx.doi.org/10.21437/SpeechProsody.2016-207. Acesso em: 06 maio 2021.

LANGARANI, M. S. E.; SANTEN, J. V.; MOHAMMADI, S. H.; KAIN, A. Data-driven foot-based intonation generator for text-to-speech synthesis. **Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH**, Jan., p. 1596-1600, 2015. Disponível em: http://dx.doi.org/10.21437/interspeech.2015-370. Acesso em: 06 maio 2021.

LATORRE, J.; LACHOWICZ, J.; LORENZO-TRUEBA, J.; MERRITT, T.; DRUGMAN, T.; RONANKI, S.; KLIMKOV, V. Effect of Data Reduction on Sequence-to-sequence Neural TTS. **ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2019, p. 7075-7079. Disponível em: https://doi.org/10.1109/ICASSP.2019.8682168. Acesso em: 22 maio 2021.

LEITE, H. M. A; CINTO, T.; CARVALHO, S. N.; PEIXOTO, C. S. A.; ARANTES, D. S. Uso de vozes sintetizadas em ambientes virtuais de ensino a distância. **Revista Ciência e Tecnologia**, v. 17, p. 1-7, 2014. Disponível em: http://www.revista.unisal.br/sj/index.php/123/article/view/344/237. Acesso em: 14 ago. 2021.

LIBERMAN, M.; PIERREHUMBERT, J. B. Intonational Invariance under Changes in Pitch Range and Length. *In*: ARONOFF, M.; OEHRLE, R. T. (eds.). **Language Sound Structure**. Cambridge: MIT Press, 1984, p. 157-233.

LIEBERMAN, P. **Intonation, Perception, and Language**. Cambridge: MIT Press, 1967.

LIU, X.; LIU, Y.; SONG, X. Investigating for Punctuation Prediction in Chinese Speech Transcriptions. 2018 International Conference on Asian Language Processing (IALP), **IEEE**, p. 74-78, 2018. Disponível em: https://doi.org/10.1109/IALP.2018.8629143. Acesso em: 25 maio 2021.

LUCENTE, L. **Aspectos dinâmicos da fala e da entoação no português brasileiro**. Tese (Doutorado em Linguística), Universidade Estadual de Campinas, Campinas, 2012.

MACIEL, D. **Os atos de fala de pedidos em português do Brasil e em alemão: um estudo interacional contrastivo**. Mestrado (Estudos da Linguagem). Universidade Federal Fluminense, Niterói, 2015.

MAIA, R.; SEARA, R. Um sistema TTS baseado em redes neurais profundas usando parâmetros síncronos de pitch. **XXXV Simpósio Brasileiro de Telecomunicações e Processamento de Sinais** – São Pedro, SP, 3-6 set., 2017. Disponível em: https://www.sbrt.org.br/sbrt2017/anais/1570361943.pdf Acesso em: 16 jul. 2021.

MANFIO, E. R. Como funcionam alguns fonemas no aplicativo Balabolka. **Revista de Linguística e Teoria Literária, Via Litterae**, Anápolis, v. 4, n. 2, p. 191-204, jul./dez. 2012. Disponível em: www2.unucseh.ueg.br/vialitterae. Acesso em: 08 ago. 2021.

MARQUES, L. B. de M. M.; UEDA, L. H.; COSTA, P. D. P. Transferência de Estilo para Síntese de Fala Expressiva. **Anais do XIV Encontro de Alunos e Docentes do DCA/FEEC/UNICAMP (EADCA)**, Campinas: UNICAMP, 2022. p. 1-6. Disponível em: https://www.dca.fee.unicamp.br/portugues/pesquisa/seminarios/2022/manuscritos/papers/18.pdf. Acesso em: 10 maio 2023.

MATEUS, M. H. M. Estudando a melodia da fala – traços prosódicos e constituintes prosódicos. **Palavras** – Revista da Associação de Professores de Português, n. 28, p. 79-98, 2005.

MATEUS, M. H. M.; FROTA, S.; VIGÁRIO, M. Aspectos fonológicos e prosódicos da gramática do português. *In*: MATEUS, M. H. M.; BRITO, A. M.; DUARTE, I.; FARIA, I. H.; VILLALVA, A. **Gramática da Língua Portuguesa**, 5a. edição revista e aumentada. Lisboa: Caminho, 2003.

MELO, M. C. **Ensino da prosódia nos atos diretivos de ordem e pedido para falantes estrangeiros aprendizes do Português Brasileiro: uma análise de materiais didáticos**. Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina, 2017. Dissertação (Mestrado Profissional em Educação). Disponível em: http://acervo.ufvjm.edu.br/jspui/bitstream/1/1655/1/maressa_carneiro_melo.pdf Acesso em: 30 ago. 2021.

MING, H.; HUANG, D.; DONG, M.; LI, H.; XIE, L.; ZHANG, S. Fundamental frequency modeling using wavelets for emotional voice conversion. **International Conference on Affective Computing and Intelligent Interaction (ACII)**, 2015, pp. 804-809, Disponível em: https://doi.org/10.1109/ACII.2015.7344665. Acesso em: 18 jul. 2021.

MIRANDA, L. S. Entoação do Português do Brasil: uma descrição perceptiva. **Anais do Congresso Brasileiro de Prosódia**, v. 2, 2013. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/anais_coloquio/article/view/6163/5337 Acesso em: 30 ago. 2021.

MIRANDA, L. S.; MORAES, J. A. A percepção de valores pragmáticos na entoação de sentenças imperativas no português brasileiro: um estudo experimental. **Diadorim**, Rio de Janeiro, v. 20 – Especial, p. 263-290, 2018. Disponível em: https://revistas.ufrj.br/index.php/diadorim/article/view/23277/13011 Acesso em: 30 ago. 2021.

MORAES, J. A.; COLAMARCO, M. Você está pedindo ou perguntando? Uma análise entonacional de pedidos e perguntas no português do Brasil. **Revista de Estudos da Linguagem**, Belo Horizonte, v. 15, n. 2, p. 113-126, jul./dez. 2007. Disponível em: http://dx.doi.org/10.17851/2237-2083.15.2.113-126. Acesso em: 10 set. 2021.

MORAES, J. A.; RILLIARD, A. Entoação. *In*: Miguel Oliveira Jr. (Org.). **Prosódia, Prosódias**. 1ed. São Paulo: Contexto, 2022, v. 1, p. 45-66.

MORAES, J. A., RILLIARD, A. MOTA, B.; SHOCHI, T. Multimodal perception and production of attitudinal meaning in Brazilian Portuguese. **Proceedings Speech Prosody**, 2010.

MOREIRA, N. A. M. **Proposta de um front-end em java para sintetizador de voz baseado no MBROLA**. Dissertação (Engenharia de Teleinformática). Universidade Federal do Ceará, Centro de Tecnologia, Departamento de Engenharia de Teleinformática, Fortaleza, 2015.

MOUNGSRI. D.; KORIYAMA, T.; KOBAYASHI, T. Enhanced F0 generation for GPR-based speech synthesis considering syllable-based prosodic features. **Asia-Pacific Signal and Information Processing Association Annual Summit and Conference** (APSIPA ASC), 2017, p. 1524-1527, Disponível em: https://doi.org/10.1109/APSIPA.2017.8282285. Acesso em: 18 jul. 2021.

MUSILIYIU, O.; OLIVEIRA JR., M. Padrões entoacionais dos números telefônicos no português brasileiro. **Revista da ABRALIN**, v. 14, n. 1, 10 ago. 2015. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1249. Acesso em: 06 jun. 2022.

NETO, N. S.; SOUZA, E.; MACEDO, V.; ADAMI, A. G.; KLAUTAU, A. Desenvolvimento de software livre usando reconhecimento e síntese de voz: o estado da arte para o português brasileiro. **ACM International Conference Proceeding Series**, 2005, p. 326-331, v. 124. Disponível em: https://doi.org/10.1145/1111360.1111396. Acesso em: 24 jan. 2022.

NOAH, B.; SETHUMADHAVAN, A.; LOVEJOY, J.; MONDELLO, D. Public perceptions towards synthetic voice technology. **Proceedings of the Human Factors and Ergonomics Society Annual Meeting**. Sage CA: Los Angeles, CA: SAGE Publications, 2021. p. 1448-1452. Disponível em: https://doi.org/10.1177/1071181321651128. Acesso em: 03 jun. 2021.

OLIVEIRA, L. M. V. V. C. **Síntese de fala a partir de texto.** Dissertação (Mestrado em Engenharia Electrotécnica e de Computadores). Universidade Técnica de Lisboa, Instituto Superior Técnico, Lisboa, 1996.

OLIVEIRA JR., M. **Prosodic features in spontaneous narratives**. Thesis (Doctor of Philosophy). Department of Linguistics, Simon Fraser University, Vancouver, 2000.

OLIVEIRA JR., M. Aspectos técnicos na coleta de dados linguísticos orais. *In*: FREITAG, R. M. K. (Ed). **Metodologia de Coleta e Manipulação de dados em Sociolinguística**. São Paulo: Blucher, 2014.

OLIVEIRA JR., M. NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). **CHIMERA. Romance Corpora and Linguistic Studies**, v. 3, n. 2, p. 149-174, 2016. Disponível em: https://doi.org/10.15366/chimera2016.3.2.004. Acesso em: 13 mar. 2023.

PACHECO, F. S. Artigo de Revisão: Sistemas de Síntese de Fala. **Revista Ilha Digital**, ISSN 2177-2649, v. 2, p. 3 – 17, 2010. Disponível em: http://ilhadigital.florianopolis.ifsc.edu.br/index.php/ilhadigital/article/view/17. Acesso em: 07 ago. 2021.

PAIVA, S.; MOUTINHO, L. C.; TEIXEIRA, A. J. S. Síntese por concatenação de variantes regionais: o falar do Porto. **Actas do XX Encontro Nacional da Associação Portuguesa de Linguística**, Lisboa, APL, 2004. Disponível em: https://apl.pt/wp-content/uploads/2017/12/2004-61.pdf. Acesso em: 08 ago. 2021.

PIERREHUMBERT, J. **The Phonology and Phonetics of English Intonation**. Ph.D thesis, MIT, 1980.

PORTET, F.; VACHER, M.; GOLANSKI, C.; ROUX, C.; MEILLON, B. Design and evaluation of a smart home voice interface for the elderly: acceptability and objection aspects. **Personal and Ubiquitous Computing**, v. 17, n. 1, p. 127-144, 2013. Disponível em: https://doi.org/10.1007/s00779-011-0470-5. Acesso em: 03 maio 2023.

PRICE, P.; OSTENDORF, M.; SHATTUCK-HUFNAGEL, S.; FONG, C. The use of prosody in syntactic disambiguation. **Journal of the Acoustic Society of America**, v. 90, n. 6, p. 2956-2970, 1991. Disponível em: https://doi.org/10.1121/1.401770. Acesso em: 03 maio 2023.

QUEIROZ, H. S. **A contribuição da prosódia e da qualidade de voz na expressão de atitudes do locutor em atos de fala diretivos**. Tese (Doutorado em Linguística) – Universidade Federal de Minas Gerais, Minas Gerais, 2011.

RAO, M.V. A.; GHOSH, P. K. Pitch prediction from Mel-generalized cepstrum — a computationally efficient pitch modeling approach for speech synthesis. **2017 25th European Signal Processing Conference (EUSIPCO)**, 2017, p. 1629-1633. Disponível em: https://doi.org/10.23919/EUSIPCO.2017.8081485. Acesso em: 22 maio 2021.

RASO, T.; ROCHA, B. Como a categoria de atitude condiciona a metodologia para o estudo das ilocuções. **Diadorim**, Rio de Janeiro, v. 2, n. 17, p. 173-197, 2015. Disponível em: https://revistas.ufrj.br/index.php/diadorim/article/viewFile/4075/3053 Acesso em: 30 ago. 2021.

R Core Team. **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2016-2023. Disponível em: https://www.r-project.org/. Acesso em: 01 maio 2023.

RE, D. E.; O'CONNOR, J. J.; BENNETT, P. J.; FEINBERG, D. R. Preferences for very low and very high voice pitch in humans. **PloS one**, v. 7, n. 3, 2012. Disponível em: https://doi.org/10.1371/journal.pone.0032719. Acesso em: 03 jun. 2023.

REED, B. S. Intonation phrases in natural conversation: a participants' category? *In*: BARTH-WEINGARTEN, D.; REBER, E.; SELTING, M. (eds.). **Prosody in Interaction**. Amsterdam: John Benjamins, 2010, p. 191-212.

REED, B. S. **Analysing Conversation**: an introduction to prosody. Houndmills: Palgrave Macmillan, 2011.

REDDY, R.; RAO, S. Prosody modeling for syllable based text-to-speech synthesis using feed forward neural networks. **Neurocomputing**, v. 171, n. 1, Jan., p. 1323-1334, 2015. Disponível em: http://dx.doi.org/10.1016/j.neucom.2015.07.053i. Acesso em: 06 maio 2021.

RIBEIRO, M. S.; YAMAGISHI, J.; CLARK, R. A. J. A perceptual investigation of wavelet-based decomposition of F0 for text-to-speech synthesis. **INTERSPEECH, 16th Annual Conference of the International Speech Communication Association, Dresden**, Germany, Sep. 6-10, 2015. Disponível em: https://www.isca-speech.org/archive/interspeech_2015/i15_1586.html. Acesso em: 23 abr. 2021.

ROBINSON, C.; OBIN, N.; ROEBEL, A. Sequence-to-sequence Modelling of F0 for Speech Emotion Conversion. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, 2019, p. 6830-6834. Disponível em: https://doi.org/10.1109/ICASSP.2019.8683865. Acesso em: 18 jul. 2021.

ROCHA, B. Metodologia empírica para o estudo de ilocuções naturais do PB. **Domínios de Lingu@gem**, v. 7, n. 2, p. 109-148, 20 dez. 2013. Disponível em: http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/23747/13573 Acesso em: 30 ago. 2021.

ROCHA, B. **Uma metodologia empírica para a identificação e descrição de ilocuções e a sua aplicação para o estudo da Ordem em PB e italiano**. Tese (Doutorado em Linguística). Universidade Federal de Minas Gerais, Belo Horizonte, 2016.

ROCHA, B. M. **A unidade informacional de introdutor locutivo no português brasileiro: uma análise baseada em corpus**. Dissertação (Mestrado em Linguística), Universidade Federal da Minas Gerais, Belo Horizonte, 2011.

RONANKI, S. **Prosody generation for text-to-speech synthesis**. Tese (Doctor Of Philosophy), Institute for Language, Cognition and Computation School of Informatics, University of Edinburgh, 2019.

SÁ, F. C. **Geração de prosódia para o português brasileiro em sistemas *text-to-speech***. Monografia (Bacharelado em Ciência da Computação). Universidade Federal do Rio Grande do Norte, Natal, 2018.

SAGISAKA, Y. Speech synthesis from text. **IEEE Communications Magazine**, v. 28, n. 1, 35–41, 1990. Disponível em: https://doi.org/10.1109/35.46669. Acesso em: 07 ago. 2021.

SAITO, T. On the use of F0 features in automatic segmentation for speech synthesis. **5th International Conference on Spoken Language Processing**. Sydney, Australia November 30 - December 4, 1998. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.28.330 Acesso em: 08 ago. 2021.

SANTEN, J. P. V.; MÖBIUS, B.; VENDITTI, J. J.; SHIH, C. Description of the Bell Labs intonation system. **The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis**, 1998. Disponível em: https://www.isca-speech.org/archive_open/archive_papers/ssw3/ssw3_293.pdf. Acesso em: 01 jun. 2023.

SANTOS, A. J. **O papel dos marcadores prosódicos na fluência de leitura**. Dissertação (Mestrado em Linguística). Universidade Estadual do Sudoeste da Bahia (UESB), Vitória da Conquista, Bahia, 2016.

SANTOS, P. S. Uma proposta de descrição prosódica dos atos de fala ordem e pedido voltada para o ensino de português como língua estrangeira (PLE). Revista Virtual de **Estudos da Linguagem**, v. 8, n. 15, 2010. Disponível em: http://www.revel.inf.br/files/artigos/revel_15_uma_proposta_de_descricao_prosodica.pdf Acesso em: 30 ago. 2021.

SCHÖTZ, S.; BESKOW, J.; BRUCE, G.; GUSTAFSON, J.; GRANSTRÖM, B.; SEGERUP, M. Synthesising intonational varieties of Swedish. **Working papers/Lund University, Department of Linguistics and Phonetics**, v. 54, p. 85-90, 2010. Disponível em: https://journals.lub.lu.se/LWPL/article/view/19651. Acesso em: 01 jun. 2010.

SCORDILIS, M. S.; GOWDY, J. N. Neural network based generation of fundamental frequency contours. **International Conference on Acoustics, Speech, and Signal Processing**, p. 219-222 v. 1, 1989. Disponível em: https://doi.org/10.1109/ICASSP.1989.266404. Acesso em: 10 ago. 2021.

SEARLE, J. R. **Os actos de fala: um ensaio de filosofia da linguagem**. Coimbra: Almedina, 1991.

SEARLE, J. R. **Expressão e significado: estudos da teoria dos atos de fala.** Tradução: Ana Cecília G. A. de Camargo e Ana Luiza M. Garcia. São Paulo: Martins Fontes, 1995.

SERRA, C. R. **Realização e percepção de fronteiras prosódicas no português do Brasil: fala espontânea e leitura.** Rio de Janeiro: 2009. Tese (Doutorado em Linguística). Faculdade de Letras, Universidade Federal do Rio Janeiro, Rio de Janeiro.

SILVA, C. H.; NAGLE, E. J.; RUNSTEIN, F.; VIOLARO, F. F0 generation in a text-to-speech system using a database of natural F0 patterns. **Proceedings. SBT/IEEE International Telecommunications Symposium**, p. 213-218 v.1, 1998. Disponível em: https://doi.org/10.1109/ITS.1998.713119. Acesso em: 08 ago. 2021.

SILVA, C. H.; VIOLARO, F. Modelamento prosódico para conversão texto-fala do português falado no Brasil. **Revista Brasileira de Telecomunicações**, v. 10, n. 1, 1995. Disponível em: https://jcis.sbrt.org.br/jcis/article/view/179/93. Acesso em: 08 ago. 2021.

SILVA, S. Z. **Um estudo de modelos básicos de prosódia para o Português Brasileiro**. Tese (Mestrado em Engenharia Elétrica), Universidade Federal do Rio de Janeiro, COPPE, Rio de Janeiro, 2004.

SIMÕES, F. O. I**mplementação de um sistema de conversão texto-fala para o português do Brasil.** Dissertação (Mestrado em Engenharia Elétrica), Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, 1999.

SIMÕES, F. O.; VIOLARO, F.; BARBOSA, P. A.; ALBANO, E. C. Um sistema de conversão texto-fala para o português falado no Brasil. **Journal of Communication and**

Information Systems, v. 15, n. 2, 2000. Disponível em: http://dx.doi.org/10.14209/jcis.2000.8. Acesso em: 08 ago. 2021.

SONNTAG, G. P.; PORTELE, T.; HEUFT, B. Prosody generation with a neural network: weighing the importance of input parameters. **IEEE International Conference on Acoustics, Speech, and Signal Processing,** p. 931-934 v.2, 1997. Disponível em: https://doi.org/10.1109/ICASSP.1997.596089. Acesso em: 08 ago. 2021.

SOUZA, C. F. S. S. **Síntese de Fala em Português Brasileiro Baseada em Modelos Ocultos de Markov**. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Pernambuco, 2010.

SVARTMAN, F. R. F. O sistema de notação ToBI. *In*: **Verbetes LBASS**. 2021. Disponível em: http://www.letras.ufmg.br/lbass/. Acesso em: 03 maio 2023.

SWERTS, M. Prosodic features at discourse boundaries of different strength. **Journal of the Acoustical Society of America**, v. 101, n. 1, p. 514-521, 1997. Disponível em: https://doi.org/10.1121/1.418114. Acesso em: 13 mar. 2023.

SZASZÁK, G., TÜNDIK, M. Á. Leveraging a character, word and prosody triplet for an ASR error robust and agglutination friendly punctuation approach. **Proc. Interspeech**, p. 2988-2992, 2019. Disponível em: http://dx.doi.org/10.21437/Interspeech.2019-2132. Acesso em: 25 maio 2021.

TAHON, M.; LECORVÉ, G.; LOLIVE, D. Can We Generate Emotional Pronunciations for Expressive Speech Synthesis? **IEEE Transactions on Affective Computing**, v. 11, n. 4, p. 684-695, 1 Oct.-Dec. 2020. Disponívl em: https://doi.org/10.1109/TAFFC.2018.2828429. Acesso em: 18 jul. 2021.

TAYLOR, P. **Text–to–Speech Synthesis**. Cambridge University Press, 2009.

TEIXEIRA, A. H. K.; SANTOS, I. M. M.; MOTA, J. S.; GOMES DE SOUZA, J. Tecnologias de reconhecimento de fala: uma revisão sistemática de trabalhos no Brasil. **XX Encoinfo – Congresso de Computação e Tecnologias da Informação**, p. 160-167, 2018. Disponível em: http://ulbra-to.br/encoinfo/wp-content/uploads/2020/03/Tecnologias-de-Reconhecimento-de-Fala-uma-revis%C3%A3o-sistem%C3%A1tica-de-trabalhos-no-Brasil.pdf. Acesso em: 13 mar. 2022.

TENANI, L. **Domínios prosódicos no Português do Brasil: implicações para a prosódia e para a aplicação de processos fonológicos.** Tese (Doutorado em Linguística). Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2002.

TENANI, L. Estruturas morfossintática e prosódica dos enunciados: fatores para hipersegmentações. **Domínios de Lingu@gem**, v. 11, n. 3, p. 600-626, jul./set. 2017. DOI: 10.14393/DL30-v11n3a2017-8. Acesso em: 20 jun. 2022.

TENANI, L.; PARANHOS, F. C. Análise prosódica de segmentações não convencionais de palavras em textos do sexto ano do Ensino Fundamental. **Filologia e Linguística Portuguesa**, v. 13, n. 2, p. 477-504, 2011. Disponível em: https://doi.org/10.11606/issn.2176-9419.v13i2p477-504. Acesso em: 20 jun. 2022.

TEVAH, R. T. **Implementação de um sistema de reconhecimento de fala contínua com amplo vocabulário para o português brasileiro.** Dissertação (Mestrado em Ciências em Engenharia Elétrica), Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2006.

THOMAS, C.; GOKUL, P.; THOMAS, N.; GOPINATH, D. P. Synthesizing intonation for Malayalam TTS. **International Conference on Control Communication & Computing India (ICCC),** 2015, p. 522-527, Disponível em: https://doi.org/10.1109/ICCC.2015.7432949. Acesso em: 22 maio 2021.

TÓTH, B. P.; CSAPÓ, T. G. Continuous fundamental frequency prediction with deep neural networks**. 24th European Signal Processing Conference (EUSIPCO)**, 2016, p. 1348-1352, Disponível em: https://doi.org/10.1109/EUSIPCO.2016.7760468. Acesso em: 18 jul. 2021.

VANDERVEKEN, D. O que é uma força ilocucional? **Cadernos de Estudos Linguísticos**. Campinas, v. 9. 1985, p. 173-194. Tradução: João Wanderley Geraldi. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8636734 Acesso em: 10 set. 2021.

VASSOLER, A. M. de O.; MEDEIROS, B. R. de. Frequência Fundamental e Emoções: um estudo a partir da fala atuada em português brasileiro. **Signum: Estudos da Linguagem**, v. 16, n. 2, p. 327-352, 2013. Disponível em: https://doi.org/10.5433/2237-4876.2013V16N2P327. Acesso em: 01 jun. 2023.

VECCHIETTI, L. F. S. **Processamento de uma nova base de voz com aplicação em síntese de fala utilizando modelos ocultos de Markov.** Projeto de Graduação (Engenharia). Universidade Federal do Rio de Janeiro, Escola Politécnica, Rio de Janeiro, 2015.

WAGNER, P.; BESKOW, J.; BETZ, S.; EDLUND, J.; GUSTAFSON, J.; EJE HENTER, G.; MAGUER, S. L.; MALISZ, Z.; SZÉKELY, É. TÅNNANDER, C.; VOßE, J. Speech synthesis evaluation—state-of-the-art assessment and suggestion for a novel research program. **Proceedings of the 10th Speech Synthesis Workshop (SSW10)**. 2019. Disponível em: http://dx.doi.org/10.21437/SSW.2019-19. Acesso em: 08 jun. 2023.

WANG, H.; SOONG, F.; MENG, H. AA spectral space warping approach to cross-lingual voice transformation in HMM-based TTS. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP***)*, 2015, p. 4874-4878. Disponível em: https://doi.org/10.1109/ICASSP.2015.7178897. Acesso em: 23 abr. 2021.

WANG, X.; TAKAKI, S.; YAMAGISHI, J. An RNN-based quantized F0 model with multi-tier feedback links for text-to-speecch synthesis. **INTERSPEECH, Computer Science**. Aug. 20-24, Stockholm, Sweden, 2017. Disponível em: https://doi.org/10.21437/Interspeech.2017-246. Acesso em: 23 abr. 2021.

WICKHAM, H. **ggplot2**: Elegant Graphics for Data Analysis. New York: Springer-Verlag, 2016. Disponível em: https://ggplot2.tidyverse.org. Acesso em: 01 maio 2023.

WIGHTMAN, C.; SHATTUCK-HUFNAGEL, S.; OSTENDORF, M.; PRICE, P. J. Segmental durations in the vicinity of prosodic phrase boundaries. **Journal of the Acoustical**

Society of America, v. 91, n. 3, p. 1707-1717, 1992. Disponível em: https://doi.org/10.1121/1.402450. Acesso em: 03 maio 2023.

XU, Y. ProsodyPro – A Tool for Large-scale Systematic Prosody Analysis. **TRASP 2013 Proceedings**, Aix-em-Provence, France, p. 7-10, 2013. Disponível em: https://www.homepages.ucl.ac.uk/~uclyyix/ProsodyPro/. Acesso em: 13 mar. 2023.

YADAV, J.; RAO, K. S. Generation of emotional speech by prosody imposition on sentence, word and syllable level fragments of neutral speech. **International Conference on Cognitive Computing and Information Processing** (CCIP*)*, 2015, p. 1-5. Disponível em: https://doi.org/10.1109/CCIP.2015.7100694. Acesso em: 18 jul. 2021.

YI, J.; TAO, J. Self-attention Based Model for Punctuation Prediction Using Word and Speech Embeddings. **IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)**, Brighton, UK, 2019, p. 7270-7274. Disponível em: https://doi.org/10.1109/ICASSP.2019.8682260. Acesso em: 08 fev. 2021.

YIN, X.; LEI, M.; QIAN, Y.; SOONG, F. K.; HE, L.; LING, Z.; DAI, L. Modeling F0 trajectories in hierarchically structured deep neural networks. **Speech Communication**, v. 76, p. 82-92, 2016. Disponível em: https://doi.org/10.1016/j.specom.2015.10.007. Acesso em: 06 maio 2021.

ZEN, H.; NOSE, T.; YAMAGISHI, J.; SAKO, S.; MASUKO, T.; BLACK, A. W.; TOKUDA, K. The HMM-based Speech Synthesis System (HTS) Version 2.0. **6th ISCA Workshop on Speech Synthesis**, Bonn, Germany, Aug. 22-24, 2007. Disponível em: http://www.cs.cmu.edu/~awb/papers/ssw6/ssw6_294.pdf. Acesso em: 16 set. 2021.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. **Speech Communication**, Elsevier, v. 51, n. 11, p. 1039–1064, 2009. Disponível em: https://doi.org/10.1016/j.specom.2009.04.004. Acesso em: 16 set. 2021.

ZHAO, Y.; LI, H.; LAI, C.; WILLIAMS, J.; COOPER, E.; YAMAGISHI, J. Improved prosody from learned F0 codebook representations for VQ-VAE speech waveform reconstruction. **Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH**, Oct., n. 1, p. 4417-4421, 2020. Disponível em: https://doi.org/10.48550/arXiv.2005.07884. Acesso em: 06 maio 2021.

APÊNDICE A – FRASES SELECIONADAS DO GOOGLE MAPS E DO WAZE

Frases do *Google Maps*
1. Siga na direção oeste, na Rua Hélio Pradines, em direção à Rua José Cabral Acioli e depois vire à direita na Rua José Cabral Acioli. 2. Vire à direita na Rua José Cabral Acioli. 3. Em 300 metros, continue em frente para permanecer na Rua José Cabral Acioli. 4. Continue em frente para permanecer na Rua José Cabral Acioli e depois vire à direita na Rua Ernesto Gomes Maranhão. 5. Vire à direita na Rua Ernesto Gomes Maranhão e depois vire à esquerda na Rua Artagnan Martins Reis. 6. Vire à esquerda na Avenida Doutor Antônio Gomes de Barros. 7. Continue na Avenida Doutor Antônio Gomes de Barros por 800 metros. 8. Em 400 metros, use a faixa da esquerda para virar à direita e acessar a Avenida Dona Constança de Góes Monteiro. 9. Vire à esquerda e depois vire à esquerda. 10. Siga na direção leste na Avenida Luís Ramalho de Castro em direção à Rua Doutor José Bernardino Carvalho Leite. 11. Vire à direita na Rua Doutor José Bernardino Carvalho Leite. 12. Em 200 metros, vire à direita na Rua Abdom Assis Inojosa Andrade. 13. Vire à direita. 14. Vire à esquerda na Rua Hamilton de Barros Soutinho e depois vire à direita na Rua Soldado Eduardo dos Santos. 15. Em 300 metros, vire à esquerda. 16. Vire à direita na Travessa Antônio Maciel de Oliveira. 17. Em 200 metros, vire à esquerda na Rua Professor Sandoval Arroxelas. 18. Em 400 metros, vire à direita na Rua Desportista Humberto Guimarães. 19. Em 500 metros, vire à direita na Rua José Júlio Sawer. 20. Vire à direita na Rua José Júlio Sawer e depois vire à esquerda na Rua Hélio Pradines.

Frases do *Waze*
1. Vire à direita em Rua José Cabral Acioli. 2. Em 500 metros, vire à direita em Rua Ernesto Gomes Maranhão. 3. Em 300 metros, vire à direita em Rua Ernesto Gomes Maranhão. Em seguida, vire à esquerda. 4. Vire à direita em Rua Ernesto Gomes Maranhão. Em seguida, vire à esquerda em Rua Artagnan Martins Reis. 5. Vire à esquerda em Avenida Doutor Antônio Gomes de Barros. 6. Em 900 metros, mantenha a esquerda. 7. Vire à direita em Rua Carlos de Lima. Em seguida, vire à direita em Rua Pretestato Ferreira Machado. 8. Vire à direita em Avenida Pretestato Ferreira Machado. 9. Em 100 metros, vire à direita. Em seguida, retorne. 10. Vire à direita em Avenida Doutor Antônio Gomes de Barros. Em seguida, vire à esquerda. 11. Vire à direita. Em seguida, vire à esquerda. 12. Em 200 metros, vire à direita em Rua Soldado Eduardo dos Santos. 13. Vire à esquerda em Avenida Professor Sandoval Arroxelas. Em seguida, vire à direita em Rua José Freire Moura. 14. Em 400 metros, vire à direita em Rua Desportista Humberto Guimarães. 15. Vire à direita em Rua Desportista Humberto Guimarães. 16. Vire à direita. 17. Em 400 metros, vire à esquerda em Avenida Professor Sandoval Arroxelas. 18. Em 300 metros, vire à esquerda. 19. Em 400 metros, vire à direita em Rua José Júlio Sawer. 20. Vire à direita em Rua José Júlio Sawer. Em seguida, vire à esquerda em Rua Hélio Pradines.

APÊNDICE B – INCLUSÃO E DE EXCLUSÃO DAS FRASES DO GPS

Frases do *Google Maps*	Frases do *Waze*
1. Siga na direção oeste, na Rua Hélio Pradines, em direção à Rua José Cabral Acioli e depois vire à direita na Rua José Cabral Acioli. (incluída, após os critérios de equivalências e por ordem de apresentação)	1. Vire à direita em Rua José Cabral Acioli. (incluída, pois é equivalente à frase 2 do *Maps*)
2. Vire à direita na Rua José Cabral Acioli. (incluída, por ser equivalente à frase 1 do *Waze*)	2. Em 500 metros, vire à direita em Rua Ernesto Gomes Maranhão. (incluída, após os critérios de equivalências e por ordem de apresentação)
3. Em 300 metros, continue em frente para permanecer na Rua José Cabral Acioli. (incluída, após os critérios de equivalências e por ordem de apresentação).	3. Em 300 metros, vire à direita em Rua Ernesto Gomes Maranhão. Em seguida, vire à esquerda. (incluída, após os critérios de equivalências e por ordem de apresentação)
4. Continue em frente para permanecer na Rua José Cabral Acioli e depois vire à direita na Rua Ernesto Gomes Maranhão. (incluída, após os critérios de equivalências e por ordem de apresentação).	4. Vire à direita em Rua Ernesto Gomes Maranhão. Em seguida, vire à esquerda em Rua Artagnan Martins Reis. (incluída, pois é equivalente à frase 5 do *Maps*)
5. Vire à direita na Rua Ernesto Gomes Maranhão e depois vire à esquerda na Rua Artagnan Martins Reis. (incluída, por ser equivalente à frase 4 do *Waze*)	5. Vire à esquerda em Avenida Doutor Antônio Gomes de Barros. (incluída, pois é equivalente à frase 7 do *Maps*)
6. Vire à esquerda na Rua Artagnan Martins Reis e depois vire à esquerda na Avenida Doutor Antônio de Barros. (excluída, pois é equivalente à frase 5).	6. Em 900 metros, mantenha a esquerda. (incluída, após os critérios de equivalências e por ordem de apresentação)
7. Vire à esquerda na Avenida Doutor Antônio Gomes de Barros. (incluída, pois é equivalente à frase 5 do *Waze*)	7. Em 300 metros, mantenha a esquerda. (excluída, pois é equivalente à frase 6)
8. Continue na Avenida Doutor Antônio Gomes de Barros por 800 metros. (incluída, após os critérios de equivalências de frases e por ordem de apresentação).	8. Mantenha a esquerda. (excluída, pois é equivalente à frase 33, uma vez que tem a mesma estrutura, diferenciando-se o léxico, que não é considerado como análise neste trabalho)
9. Em 400 metros, use a faixa da esquerda para virar à direita e acessar a Avenida Dona Constança de Góes Monteiro. (incluída, após os critérios de equivalências e por ordem de apresentação).	9. Vire à direita em Rua Carlos de Lima. Em seguida, vire à direita em Rua Pretestato Ferreira Machado. (incluída, após os critérios de equivalências e por ordem de apresentação)
10. Use a faixa da esquerda para virar à direita e acessar a Avenida Dona Constança de Góes Monteiro e depois vire à esquerda. (excluída, por ser equivalente à frase 9)	10. Vire à direita em Avenida Pretestato Ferreira Machado. (incluída, pois é equivalente à frase 14 do *Maps*)
11. Vire à esquerda e depois vire à esquerda. (incluída, pois é equivalente à frase 14 do *Waze*)	11. Em 100 metros, vire à direita. Em seguida, retorne. (incluída, após os critérios de equivalências e por ordem de apresentação)
12. Vire à esquerda. (excluída, pois a diferença entre essa frase e a frase 20 é o léxico, o que não vai ser considerado como análise neste trabalho)	12. Vire à esquerda em Avenida Paulo Falcão. (excluída, por ser equivalente à frase 1)
	13. Vire à direita em Avenida Doutor Antônio Gomes de Barros. Em seguida, vire à esquerda.

13. Siga na direção leste na Avenida Luís Ramalho de Castro em direção à Rua Doutor José Bernardino Carvalho Leite. (incluída, após os critérios de equivalências e por ordem de apresentação)	(incluída, após os critérios de equivalências e por ordem de apresentação) 14. Vire à direita. Em seguida, vire à esquerda. (incluída, pois é equivalente à frase 11 do *Maps*)
14. Vire à direita na Rua Doutor José Bernardino Carvalho Leite. (incluída, pois é equivalente à frase 10 do *Waze*)	15. Vire à esquerda em Rua Júlio de Almeida Braga. Em seguida, vire à esquerda em Rua Ernesto Gomes Maranhão. (excluída, pois é equivalente à frase 9)
15. Em 200 metros, vire à direita na Rua Abdom Assis Inojosa Andrade. (incluída, pois é equivalente à frase 18 do *Waze*)	16. Vire à esquerda em Rua Ernesto Gomes Maranhão. Em seguida, vire à direita em Rua Soldado José Guilherme da Silva. (excluída, por ser equivalente à frase 9)
16. Vire à direita na Rua Abdom Assis Inojosa Andrade e depois vire à esquerda na Rua Luísa Medeiros. (excluída, pois é equivalente à frase 5)	17. Vire à esquerda em Rua Hamilton de Barros Soutinho. (excluída, por ser equivalente à frase 1)
17. Vire à esquerda, depois, vire à esquerda. (excluída, pois é equivalente à frase 11)	18. Em 200 metros, vire à direita em Rua Soldado Eduardo dos Santos. (incluída, pois é equivalente à frase 15 do *Maps*)
18. Vire à esquerda. (excluída, pois a diferença entre essa frase e a frase 20 é o léxico, o que não vai ser considerado como análise neste trabalho)	19. Vire à direita em Rua Soldado Eduardo dos Santos. (excluída, por ser equivalente à frase 1)
19. Depois... (excluída, pois não é uma frase diretiva)	20. Em 200 metros, vire à esquerda em Rua Deputado José Lages. (excluída, por ser equivalente à frase 2)
20. Vire à direita. (incluída, pois é equivalente à frase 33 do *Waze*).	21. Vire à esquerda em Rua Deputado José Lages. (excluída, por ser equivalente à frase 1)
21. Vire à direita, depois, vire à esquerda. (excluída, pois é equivalente à frase 11)	22. Você chegou. Destino está à sua direita. (excluída, pois não é uma frase diretiva)
22. Vire à esquerda na Rua Hamilton de Barros Soutinho e depois vire à direita na Rua Soldado Eduardo dos Santos. (incluída, pois é equivalente à frase 26 do *Waze*)	23. Em 200 metros, vire à direita em Travessa Antônio Maciel de Oliveira. (excluída, por ser equivalente à frase 2)
23. Vire à direita. (excluída, pois é igual à frase 20)	24. Vire à direita em Travessa Antônio Maciel de Oliveira. (excluída, pois é equivalente à frase 1)
24. Em 300 metros, vire à esquerda. (incluída, após os critérios de equivalências e por ordem de apresentação).	25. Em 200 metros, vire à esquerda em Avenida Professor Sandoval Arroxelas. Em seguida, vire à direita. (excluída, pois é equivalente à frase 3)
25. Vire à esquerda na Rua Deputado José Lages e depois seu destino estará à direita. (excluída, pois é equivalente à frase 5)	26. Vire à esquerda em Avenida Professor Sandoval Arroxelas. Em seguida, vire à direita em Rua José Freire Moura. (incluída, pois é equivalente à frase 22 do *Maps*)
26. Em 600 metros, vire à esquerda. (excluída, por ser equivalente à frase 24)	27. Vire à direita em Rua José Freire Moura. (excluída, pois é equivalente à frase 1)
27. Siga na direção Leste na Rua Deputado José Lages em direção à Rua José Júlio Sawer. (excluída, após os critérios de equivalências e por ordem de apresentação)	

28. Em 150 metros, vire à direita na Travessa Antônio Maciel de Oliveira. (excluída, após os critérios de equivalências e por ordem de apresentação)	28. Em 200 metros, vire à direita em Rua Valdo Omena. (excluída, por ser equivalente à frase 2)
29. Vire à direita na Travessa Antônio Maciel de Oliveira. (incluída, pois é equivalente à frase 31 do *Waze*)	29. Vire à direita em Rua Valdo Omena. (excluída, pois é equivalente à frase 1)
30. Em 200 metros, vire à esquerda na Rua Professor Sandoval Arroxelas. (incluída, pois é equivalente à frase 34 do *Waze*)	30. Em 400 metros, vire à direita em Rua Desportista Humberto Guimarães. (incluída, pois é equivalente à frase 34 do *Maps*)
31. Vire à esquerda. (excluída, pois a diferença entre essa frase e a frase 20 é o léxico, o que não vai ser considerado como análise neste trabalho)	31. Vire à direita em Rua Desportista Humberto Guimarães. (incluída, pois é equivalente à frase 29 do *Maps*)
32. Em 500 metros, vire à direita na Rua Valdo Omena. (excluída, após os critérios de equivalências e por ordem de apresentação)	32. Você chegou. Destino está à sua direita. (excluída, pois não é uma frase diretiva)
33. Vire à direita. (excluída, pois é igual à frase 20)	33. Vire à direita. (incluída, pois é equivalente à frase 20 do *Maps*).
34. Em 400 metros, vire à direita na Rua Desportista Humberto Guimarães. (incluída, pois é equivalente à frase 30 do *Waze*)	34. Em 400 metros, vire à esquerda em Avenida Professor Sandoval Arroxelas. (incluída, pois é equivalente à frase 30 do *Maps*)
35. Vire à direita na Rua Desportista Humberto Guimarães e depois seu destino estará à direita. (excluída, após os critérios de equivalências e por ordem de apresentação)	35. Em 300 metros, vire à esquerda. (excluída, pois é equivalente à frase 9)
36. Seu destino está à direita. (excluída, pois não é uma frase diretiva)	36. Vire à esquerda em Avenida Professor Sandoval Arroxelas. (excluída, pois é equivalente à frase 34)
37. Siga para o oeste, depois, vire à direita. (excluída, pois é equivalente à frase 11)	37. Em 400 metros, vire à direita em Rua José Júlio Sawer. (incluída, pois é equivalente à frase 41 do *Maps*)
38. Vire à direita na Rua Machado Lemos. (excluída, após os critérios de equivalências e por ordem de apresentação)	38. Em 300 metros, vire à direita. Em seguida, vire à esquerda. (excluída, pois é equivalente à frase 11)
39. Em 400 metros, vire à esquerda na Rua Professor Sandoval Arroxelas. (excluída, pois é equivalente à frase 30)	39. Vire à direita em Rua José Júlio Sawer. Em seguida, vire à esquerda em Rua Hélio Pradines. (incluída, pois é equivalente à frase 42 do *Maps*)
40. Vire à esquerda na Rua Professor Sandoval Arroxelas. (excluída, pois é equivalente à frase 30)	40. Vire à esquerda em Rua Hélio Pradines. (excluída, pois é equivalente à frase 1)
41. Em 500 metros, vire à direita na Rua José Júlio Sawer. (incluída, pois é equivalente à frase 37 do *Waze*)	41. Você chegou ao seu destino. (excluída, pois não é uma frase diretiva)
42. Vire à direita na Rua José Júlio Sawer e depois vire à esquerda na Rua Hélio Pradines. (incluída, pois é equivalente à frase 39 do *Waze*)	
43. Vire à esquerda na rua Hélio Pradines e depois seu destino estará à direita.	

(excluída, após os critérios de equivalências e por ordem de apresentação) 44. Você chegou em casa. (excluída, pois não é uma frase diretiva)	

APÊNDICE C – SELEÇÃO DAS FRASES PRODUZIDAS PELAS PARTICIPANTES

Frase do GPS	Participante escolhida
Frase 1 do Google Maps	Participante 2
Frase 2 do Google Maps	Participante 1
Frase 3 do Google Maps	Participante 1
Frase 4 do Google Maps	Participante 2
Frase 5 do Google Maps	Participante 2
Frase 6 do Google Maps	Participante 1
Frase 7 do Google Maps	Participante 2
Frase 8 do Google Maps	Participante 1
Frase 9 do Google Maps	Participante 2
Frase 10 do Google Maps	Participante 2
Frase 11 do Google Maps	Participante 1
Frase 12 do Google Maps	Participante 1
Frase 13 do Google Maps	Participante 1
Frase 14 do Google Maps	Participante 2
Frase 15 do Google Maps	Participante 1
Frase 16 do Google Maps	Participante 2
Frase 17 do Google Maps	Participante 2
Frase 18 do Google Maps	Participante 2
Frase 19 do Google Maps	Participante 1
Frase 20 do Google Maps	Participante 1

Frase do GPS	Participante escolhida
Frase 1 do Waze	Participante 2
Frase 2 do Waze	Participante 1
Frase 3 do Waze	Participante 2
Frase 4 do Waze	Participante 2
Frase 5 do Waze	Participante 1
Frase 6 do Waze	Participante 1
Frase 7 do Waze	Participante 1
Frase 8 do Waze	Participante 2
Frase 9 do Waze	Participante 2
Frase 10 do Waze	Participante 1
Frase 11 do Waze	Participante 2
Frase 12 do Waze	Participante 2
Frase 13 do Waze	Participante 1
Frase 14 do Waze	Participante 1
Frase 15 do Waze	Participante 2
Frase 16 do Waze	Participante 1
Frase 17 do Waze	Participante 2
Frase 18 do Waze	Participante 1
Frase 19 do Waze	Participante 1
Frase 20 do Waze	Participante 2