MARCOS HENRIQUE CURTALE SERAFIM

Criação de um dataset de voz com múltiplos locutores para o Português Brasileiro

Trabalho de conclusão de curso de graduação apresentado como requisito para obtenção do grau de Bacharel em Ciência da Computação da Universidade Federal da Fronteira Sul.

Orientador: Prof. Dr. Fernando Bevilacqua
Este trabalho de conclusão de curso foi defendido e aprovado pela banca em: 11/05/2021

BANCA EXAMINADORA:

Dr. Fernando Bevilacqua - UFFS

Dr. Denio Duarte - UFFS

Dr. Gian Carlo Dondoni Salton - UFFS *

* remoto

Criação de um dataset de voz com múltiplos locutores para o Português Brasileiro

Marcos Henrique Curtale Serafim

[1]Universidade Federal da Fronteira Sul
Rodovia SC 484 Km 02, Bairro Fronteira Sul (Campus Chapecó) - 89.815-899 - Chapecó - SC

[2]

marcos.serafim@estudante.uffs.edu.br

Abstract. *Speech synthesis have many applications in different areas, however, the creation of good models requires a significant sample of good quality data in the language selected for the work. This work presents the creation of a public voice dataset for the Brazilian Portuguese language. A group of volunteers ($N = 29$, 10 women, 19 men) participated in the process, whose voices were recorded while they read a set of predefined sentences. The process resulted in 5800 sentences recorded which is approximately equivalent to 15 hours of raw audio. All samples were recorded under the same acoustic conditions, i.e. acoustic-insulated room with professional grade audio equipment. Our main contribution is a publicly available dataset focused on the Brazilian Portuguese language whose samples present lower or no noise artifacts in the recordings and a significantly high quality in the captured audio. The dataset can be used by researcher or practitioners interested on voice models, particularly artificial speech synthesis systems.*

1 Introdução

Sintetização de fala é uma designação de tecnologia que envolve disciplinas como acústica, linguística, processamento de sinal digital e estatística, cujo objetivo principal é a conversão de uma entrada de texto em uma saída de voz em áudio [1]. Estes sistemas estão presentes em diversas aplicações utilizadas no cotidiano, tais como assistentes virtuais, tradutores e áudio livros.

Existem diferentes métodos que realizam o processo de sintetização de fala, variando desde o concatenativo, que consiste na concatenização de segmentos de falas, até a síntese por domínio em que trechos de áudio sobre o mesmo assunto são gravados e concatenados conforme a necessidade [2]. Outra abordagem é a utilização de modelos de aprendizado mais avançados, treinados a partir de conjuntos de áudios gravados [1].

Os métodos tradicionais são suficientes para aplicações em que apenas um vocabulário limitado é requerido [2]. Para sistemas que requerem variações na voz sintetizada são necessárias diferentes abordagens [1]. Cada abordagem possui suas vantagens e desvantagens devido à dificuldade destes modelos em sintetizarem as vozes com naturalidade, ou seja, se assemelhando à forma de humanos falarem, além de outros desafios como lidar com a ambiguidade da fala, a fala espontânea e a expressão de emoções [2]. Independentemente da abordagem escolhida, o conjunto de dados utilizado no desenvolvimento deve ser na língua de saída do sistema. Além disso, as amostras devem possuir boa qualidade de áudio para garantir uma boa qualidade na saída [3].

Pesquisas no Brasil sobre sintetização de fala artificial[3, 4] demonstram que aplicações de sintetização de fala requerem uma quantidade elevada de sentenças gravadas em áudio para realizar o treinamento destes modelos. Esse número pode crescer dependendo do tipo de aplicação, que pode requerer uma ou mais vozes diferentes para realizar o aprendizado. De forma geral, quanto mais dados disponíveis para a etapa de aprendizado, melhor o modelo criado tende a ser [3]. Entretanto existem fatores que dificultam a criação destes conjuntos de dados. O custo, seja financeiro ou logístico para realizar o processo de gravação das amostras, e a própria busca por pessoas dispostas a concederem seu tempo e sua voz para serem gravadas são elementos significativos.

Atualmente, como em muitas áreas da ciência, a maior parte da produção científica é realizada na língua inglesa. Para o desenvolvimento de aplicações de aprendizado de máquina, os bancos de fala disponíveis também são, em sua maioria, disponibilizados em inglês. Tratam-se de conjuntos de dados já consolidados e utilizados amplamente em pesquisas e aplicações voltadas para o mercado [3]. Entretanto, no Brasil, trabalhos correlatos à área de modelos de fala tem apontado dificuldade em encontrar bancos de dados públicos disponíveis para pesquisa [5, 3, 4]. Esses conjuntos de dados existem, entretanto apresentam problemas. Dentre eles, estão áudios gravados com baixa qualidade, inexistência da transcrição dos textos falados e a presença de ruídos sonoros que atrapalham no aprendizado dos modelos. Esses problemas limitam a pesquisa de sintetização de fala artificial no contexto da Língua Portuguesa do Brasil, causando escassez de opções e, por vezes, forçando alternativas dentre aquelas disponíveis que não são as mais adequadas para o trabalho que se deseja realizar. A limitação de opções de conjunto de dados gera, também, um efeito colateral que é o próprio atraso na pesquisa nacional.

Com o atraso e a dificuldade na pesquisa, o cenário brasileiro de inovação e empreendedorismo é prejudicado. No lugar de novas aplicações e soluções serem criadas, fomentando novos negócios e novas parcerias, pesquisadores e entusiastas se depararam com empecilhos básicos como o acesso à informação. Dada essa carência,

este trabalho propõe a construção de um banco de dados de fala público para o idioma Português Brasileiro com múltiplos locutores, visando a disponibilizacão de áudios com qualidade em suas amostras e consistência nos seus dados e transcrições. A garantia da qualidade das amostras permite a criação de modelos de voz mais precisos, aspecto importante para a síntese de fala. A existência de múltiplos locutores permite a exploração de possibilidades de alteração na identidade de voz dos modelos de fala, bem como estudos da correlação da fala com os metadados dos locutores. Por fim, a disponibilização pública do banco de fala permite sua utilização por instituições de ensino superior para o desenvolvimento de aplicações, tal como assistentes virtuais. Permite também que pesquisadores e empresas utilizem essa base de dados para pesquisas e estudos, facilitando a realização de novos trabalhos e promovendo a inovação nacional.

2 Trabalhos relacionados

Essa secção apresenta os trabalhos referenciais para a concepção da base de dados proposta. Os trabalhos estão distribuídos nos temas de registro e documentação histórica da língua portuguesa, reconhecimento de fala, construção e processamento de bancos de fala para sintetização de fala artificial. Desta forma é possível compreender o panorama da área de processamento de linguagem natural e o contexto nacional para o desenvolvimento de pesquisas nessa área.

2.1 Norma Urbana Linguística Culta (NURC)

O Projeto da Norma Urbana Linguística Culta (Projeto NURC) foi iniciado em 1969 com a proposta inicial de documentar e estudar a norma falada culta de cinco capitais brasileiras: Recife, Salvador, Rio de Janeiro, São Paulo e Porto Alegre [6]. Como parte do material recolhido estão o registro de áudio em fitas magnéticas de entrevistas, bem como suas transcrições em texto. Os dados pertencentes ao acervo do projeto NURC foram utilizados para elaboração de trabalhos acadêmicos, incluindo dissertações de mestrado, teses de doutorado, artigos publicados em periódicos nacionais e estrangeiros, e trabalhos apresentados em encontros científicos internacionais. A maior parte destes trabalhos, entretanto, foram desenvolvidos baseados nas transcrições de texto ao invés dos arquivos de áudio. Isso foi resultado da dificuldade de acesso aos arquivos de áudio originais, armazenados em fitas magnéticas. Após a exposição de risco e perda dos dados registrados, em virtude de deterioração do material, em 2012 iniciou-se o projeto NURC Digital. Dentre os objetivos específicos do projeto está a digitalização de todo o acervo do Projeto NURC/Recife [7]. Atualmente os arquivos de áudio existem em formato digital e podem ser acessados publicamente nos portais do projeto de cada capital [6, 8, 9].

Os registros do projeto NURC foram realizados em condições variadas. Em geral as gravações foram conduzidas com microfones dinâmicos omnidirecionais, apoiados em uma mesa. Há diferentes tipos de salas utilizadas para gravação, como salas específicas, salas de aula, auditórios e, em alguns casos, nas casas dos próprios informantes. Portanto, a qualidade acústica das gravações do Projeto NURC é significativamente heterogênea [7]. Essa heterogeneidade dos registros do projeto NURC força uma seleção de amostras afim de refinar a qualidade dos dados de acordo com o objetivo do trabalho. Para pesquisas voltadas ao processamento de linguagem natural, por exemplo, não existem seleções até o momento.

2.2 Reconhecimento de fala aplicado ao Português Brasileiro utilizando *Deep Learning*

Quintanilha [5] explana sobre a problemática de encontrar bancos de voz de qualidade para o Português Brasileiro. A proposta de seu trabalho é a aplicação de um sistema de reconhecimento de fala para o Português Brasileiro utilizando Deep Learning. Esse trabalho requeriu a utilização de uma base de dados de áudio em português brasileiro para realizar o treinamento de seu modelo para posteriormente comparar com a qualidade de modelos treinados em bases de áudio em inglês.

A base de dados em inglês utilizada pelo autor foi a chamada TIMIT dataset [10], uma base de dados pela Texas Instruments e MIT com apoio financeiro da DARPA (Defense Advanced Research Projects Agency) no final de 1980. Para a base de dados em Português Brasileiro, o autor teve dificuldade de encontrar uma base de dados que atendesse a todas suas necessidades, assim o trabalho emprega seu próprio dataset a partir de outros quatro, sendo eles Spoltech Brazilian Portuguese dataset (CSLU), SID, VoxForge e LapsBM. A Tabela 1 apresenta informações relevantes que caracterizam as dimensões desses datasets. Na primeira coluna (Distribution) consta o tipo de distribuição (gratuíto ou pago), na segunda (Speakers) o número de falantes diferentes, na terceira (Utterances) o número total de amostras de áudio, na quarta (WRD/ PHN) o tipo de transcrição dos áudios. Neste caso, WRD aponta que as transcrições são a nível de palavras, equanto PHN indica que são a nível de fonemas. Por fim a quinta coluna (CE?) indica se as amostras foram gravadas em ambiente controlado.

Todos esses datasets apresentaram problemas que geraram empecilhos. Destes quatro, apenas o primeiro é privado. Os datasets CSLU e SID possuem erros de transcrição, inclusive com transcrições ausentes em alguns casos. O CSLU ainda possui muitos fonemas vocálicos com pouca frequência. O dataset VoxForge possui áudios gravados com diferentes qualidades, muitos deles com baixa qualidade. A principal característica de todos esses datasets é que suas amostras foram gravadas em ambiente não controlado, em contraste ao banco de fala em inglês (TIMIT dataset) cuja as amostras foram gravadas em ambiente controlado.

2.3 Bancos de Fala para o Português Brasileiro

Serrani and Uebel [3] apresentam 3 datasets de voz voltados para aplicações de sintetização e reconhecimento de fala para o Português Brasileiro. Cada dataset criado

Tabela 1: Datasets de fala para o português brasileiro. Reproduzido de Quintanilha [5].

Dataset	Distribution	Speakers	Utterances	WRD/PHN	CE?
CSLU: Spoltech Brazilian Portuguese	Paid	477	8.080	Both	No
Sid	Free	72	5.777	WRD	No
VoxForge	Free	+111	4.090	WRD	No
LapsBM1.4	Free	35	700	WRD	No

possui suas próprias dimensões e sentenças utilizadas no processo de gravação. O primeiro dataset, denominado como ASR-DB1, possui 248 locutores e 55.552 amostras de áudio gravadas para treinamento de algoritmos de reconhecimento de fala. Foram selecionadas 200 sentenças a partir do projeto de Alcaim [11] e adicionadas outras 24 para cobrir números cardinais e ordinais, direções, comandos, meses do ano e nomes do zodíaco. As amostras ainda possuem 16 bits de resolução e a taxa de amostragem de 48kHz. O tempo de gravação de cada locutor variou de 20 minutos até 2 horas, com média de 25 minutos por pessoa.

O segundo dataset, denominado ASR-DB2, também foi construído para aplicações de reconhecimento de fala. Possui 1.226 locutores e 815.290 amostras de áudio. Foram selecionadas 665 sentenças diferentes coletadas a partir de revistas, jornais, livros e notícias encontradas na Internet. Para a seleção das sentenças,desenvolveu-se um algoritmo que busca a melhor cobertura fonética para a língua. As amostras foram gravadas com resolução de 24 bits e 96kHz de taxa de amostragem. O tempo de gravação variou de 55 minutos até 3 horas, com média entre 1h10min e 1h30min de gravação.

O terceiro dataset, denominado TTS-DB1, foi construído para aplicações de síntese de fala. Possui 1220 sentenças selecionadas a partir do mesmo conjunto de revistas, jornais, livros e notícias com a utilização do algortimo para seleção de sentenças. As amostras foram gravadas com resolução de 24 bits e 96kHz de taxa de amostragem. No trabalho não é informado o número de locutores gravados. A Tabela 2 mostra a relação das características dos 3 datasets.

O trabalho é notavelmente abrangente. A cobertura fonética é vasta, cobrindo a língua como um todo, considerando os dialetos regionais e o estrangeirismos da língua. Essa cobertura fonética é devida ao aprimorado método de seleção de sentenças e pela distribuição dos locais de gravação nos estados da federação. Entretanto, mesmo o trabalho sendo notável, esses datasets não estão disponíveis publicamente para utilização. Pode-se especular que a causa disso possivelmente seja a parceria com a iniciativa privada, porém nenhum detalhe sobre disponibilização dos datasets é mencionado.

2.4 Base de Voz com aplicação em síntese de fala

Vecchietti [4] apresenta uma base de voz para o Português Brasileiro focada na síntese de fala, validada com posterior utilização desses dados em uma aplicação baseada em modelos ocultos de Markov. Antes de apresentar o desenvolvimento da sua própria base de dados, mencionam-se outras três. A primeira delas, a base Spoltech, contém amostras gravadas com microfone e placas de som comuns. O dataset desenvolvido no trabalho "A Brazilian Portuguese Speech Database", no qual a constituição deste é feita de forma colaborativa pelos participantes do projeto. Por fim, o corpus da Constituição Brasileira, disponibilizado pelo grupo Fala Brasil da Universidade Federal do Pará.

A base de dados final criada contém 11.456 sentenças lidas por um único locutor do sexo masculino. Consequentemente, este dataset contém apenas uma identidade de voz. As sentenças lidas foram obtidas de forma massiva a partir de livros, conteúdo jornalístico, poemas, cordéis, leituras feitas e atuadas pelo locutor simulando os sentimentos de alegria, raiva e tristeza. Há, tambem, conteúdo com expressão em frases interrogativas e exclamativas, conteúdo com estrangeirismos, comandos de voz e números. A proposta do autor é prever todas as especificidades das diversas aplicações que um sistema de síntese de fala pode ter.

A gravação foi realizada em um ambiente controlado utilizando equipamento profissional, mas não há especificação da resolução e da taxa de amostragem. O autor justifica a escolha da voz masculina como sendo a que se adequa melhor para a síntese de fala, além de critérios de disponibilidade, pontualidade, boa dicção e capacidade para trabalhar em equipe. Embora os argumentos sejam plausíveis, a justificativa da escolha da pessoal ideal para ser a locutora (masculino) não é sustentável. Por fim, o autor utiliza esses dados em modelos de Markov, porém não fornece informações em relação à disponibilização deste dataset.

2.5 Wavenet: Um modelo generativo para áudio em RAW

van den Oord [12] apresenta um dos primeiros grandes projetos de síntese de fala utilizando aprendizado de máquina para sua consolidação. As contribuições desse trabalho elucidam e guiam processos de sintetização de fala, inclusive aqueles utilizados na construção do dataset proposto no presente trabalho. O dataset utilizado nessa pesquisa foi o CSTR VCTK Corpus: *English Multi-speaker Corpus*, um corpus em inglês desenvolvido na universidade de Edinburgh. Este corpus contém amostras gravadas por 109 falantes nativos do inglês com vários sotaques.

Cada locutor gravou 400 sentenças, totalizando assim 43600 amostras. As sentenças foram selecionadas em sua maior parte de uma passagem de texto chamada Rainbow Passage [13] e a partir de notícias retiradas do jor-

Tabela 2: Bancos de fala para o Português Brasileiro

Dataset	Propósito	Sentença	N° Locutores	N° Amostras	Resolução	Taxa amostragem
ASR-DB1	Reconhecimento	224	248	55.552	16 bits	48 kHz
ASR-DB2	Reconhecimento	665	1226	815.290	24 bits	96 kHz
TTS-DB1	Síntese	1220	-	-	24 bits	96 kHz

nal The Herald (Glasgow). Cada locutor leu sentenças diferentes selecionadas a partir de um algortimo guloso projetado para maximizar a cobertura fonética. As amostras foram gravadas em um ambiente controlado, com taxa de amostragem de 96 kHz e resolução de 24 bits.

3 Metodologia

Para a construção do banco de fala proposto nesse trabalho, um processo estruturado foi elaborado, testado e replicado para o maior número possível de voluntários. Os voluntários deram seu consentimento para participar na atividade, e foram informados que poderiam desistir a qualquer momento e ter seus dados removidos. O processo ocorreu no Laboratório de Estudos Linguísticos da Universidade Federal da Fronteira Sul (UFFS), campus Chapecó. A sala possui uma acústica isolada, o que confere a diminuição significativa de ruídos para as amostras gravadas. O equipamento de captura foi um Microfone SM-58 e uma Mesa de Som Yamaha MG16XU. As amostras foram gravadas com taxa de amostragem de 44.1kHz e resolução de 32 bits. O software de gravação utilizado foi o Audacity.

Cada locutor voluntário foi submetido ao processo de gravação e coleta de metadados com duração média de 30 a 40 minutos. Ao todo, foram duzentas sentenças distintas gravadas por participante. O objetivo foi realizar a gravação de todas as sentenças para todos os falantes com a melhor qualidade possível. Por qualidade de gravação, entende-se a captura da voz na forma mais fiel possível ao som analógico, livre de ruídos e de problemas acústicos. A garantia dessa qualidade depende majoritariamente do processo de captação, sendo poucos os casos em que é possível uma melhora da qualidade do arquivo de áudio com a utilização de processos digitais [14].

O processo de gravação consistiu em duas partes principais. Primeiramente, um pesquisador recepcionou o voluntário, passando orientações gerais sobre o processo. Nesse momento também foi realizada a coleta dos metadados do voluntário, sendo eles peso, altura, etnia, sexo e idade. Na segunda parte, houve a condução do processo de gravação. Desenvolveu-se um software que guiou o voluntário na leitura das duzentas sentenças a partir do momento que ele é deixado isolado na sala de gravação. A partir do momento que é iniciada a gravação, todo o áudio é persistido em um arquivo único. O software desenvolvido é responsável pela marcação dos tempos de início e fim de cada sentença para posterior edição e corte do arquivo original.

Para a coleta dos metadados foi utilizada uma balança digital e uma fita métrica, além de um questionário integrado ao software de condução do voluntário no processo de gravação das frases. O protocolo de gravação envolveu a recepção e repasse das orientações gerais, com solicitação para remoção de calçados, casacos e itens dos bolsos antes da coleta de peso e de altura. Além desses dois dados, solicitou-se também o preenchimento pelo voluntário de um questionário demográfico informando a idade, sexo biológico e etnia. Na etapa de orientações gerais, respondeu-se eventuais dúvidas dos voluntários sobre o trabalho e o processo. Para garantir a qualidade das amostras, os voluntários foram orientados a ler as sentenças da forma mais natural e correta possível, atentando-se para a leitura fiel à forma escrita das sentenças, evitando abreviações ou cortar palavras. Voluntários também receberam orientações que poderia repetir a gravação de qualquer sentença que julgassem necessário, bastando utilizar as instruções em tela (botão para nova gravação).

Ao fim do processo de gravação, o arquivo de áudio bruto e o banco de dados do software foram persistidos em hardware e uma cópia de seguranaça foi feita. Cada voluntário possui um identificador único e anônimo que foi utilizado para a disponibilização final do banco de fala, relacionando as sentenças lidas com as respostas do questionário.

3.1 Seleção de frases para composição do banco de fala

O desenvolvimento de aplicações de sintetização de fala requer uma cobertura significativa dos fonemas da língua em que se trabalha. Os fonemas vocálicos são a menor unidade sonora de uma língua, sendo que suas uniões formam a construção dos sons das letras, sílabas e palavras. Por fim, para que esta cobertura seja significativa, é importante que os textos falados e gravados contenham amostras suficientes destes fonemas. Existe ainda o grupo de fonemas pertencentes a uma língua de forma geral, e os fonemas regionais que caracterizam os diferentes dialetos de uma língua. Como o objetivo deste trabalho é a construção de um dataset de fala para o Português Brasileiro de forma geral, os dialetos regionais não foram considerados.

As sentenças lidas pelos voluntários foram selecionadas a partir da dissertação de mestrado: Estudo Estatístico Dos Fonemas Do Português Falado Na Capital de Santa Catarina Para Elaboração de Frases Foneticamente Balanceadas [15]. Este trabalho apresenta um estudo estatístico dos fonemas do português falado na capital do estado de Santa Catarina (Florianópolis). Esse estudo abrangeu a frequência relativa média de 35 fonemas, além das frequências dos padrões silábicos do português, sílabas tônicas e não-tônicas e dos vocábulos monossílabos, dissílabos, trissílabos e polissílabos. A partir

desse estudo os autores reconhecem que há indícios que essas características podem ser generalizadas para o português falado em todo o Brasil.

Tanto o trabalho de Seara [15] quanto de outros mencionados na literatura baseiam-se nos estudos linguísticos feitos por Alcaim [11], que provê uma lista de duzentas frases foneticamente balanceadas. O trabalho de Alcaim também serviu de base para o desenvolvimento do banco de fala proposto no trabalho de Serrani and Uebel [3] e no banco de fala proposto no presente trabalho. Seara [15] apresenta como principal contribuição a construção de 20 listas de frases foneticamente balanceadas, totalizando um total de 200 frases distinas. Essas frases foram as utilizadas para o processo de gravação do presente trabalho.

3.2 Seleção de metadados dos voluntários

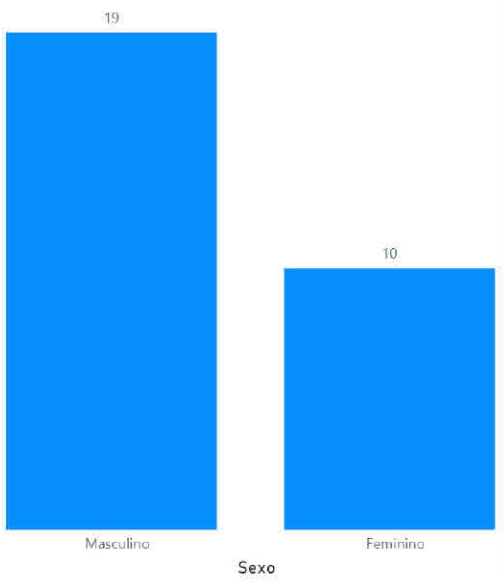
Durante a concepção do presente trabalho umas das hipóteses levantadas foi a possibilidade de geração de fala artificial modulando a identidade de voz a partir de características físicas de pessoas. Durante o processo de levantamento bibliográfico, constatou-se dificuldade de acesso a bancos de fala públicos para o Português Brasileiro. Vinsando contribuir também com essa lacuna de pesquisa, e eventualmente com trabalhos futuros, definiu-se que além da coleta de amostras de voz também seriam coletadas informações de características físicas dos voluntários. O livro Foundations of Voice Studies [16] contém um capítulo inteiramente dedicado à análise da relação entre identidade de voz e as características físicas dos falantes. Conforme mencionado, o som particular de cada locutor depende de características de sua anatomia produtora de som. Em consequência, à medida que as características físicas mudam com a maturação e envelhecimento, o som emitido também muda. As diferenças físicas entre diferentes pessoas também podem ser refletidas em diferenças consistentes em suas vozes.

Segundo a autora, existe um interesse inerente na literatura de voz sobre o que os ouvintes podem julgar a partir do som da voz. Compreender quais mudanças físicas são perceptualmente importantes e quais não produzem mudanças perceptíveis na voz de uma pessoa pode fornecer uma visão sobre as relações sociais de muitos tipos. Para a área de sintetização de fala artificial, em particular, essa informação pode contribuir para a criação de modelos de voz mais convincentes em ambientes de simulação onde a voz possui um avatar. O capítulo [16] também descreve a percepção da idade de um falante, características sexuais, origem racial ou étnica e aparência (altura, peso e características faciais) a partir da voz. Cada secção revisa as diferenças físicas que fundamentam quaisquer diferenças perceptíveis. Em seguida, examina-se os tipos de julgamentos que os ouvintes podem fazer e a maneira como extraem informações pessoais sobre os falantes das vozes. Baseado neste capítulo, definiu-se que os metadados coletados dos voluntários são a idade, sexo, etnia, altura e peso.

4 Resultados

O processo de coleta de dados com voluntários ocorreu durante os meses de Março e Abril de 2021. Em decorrência da situação global causada pela pandemia do Covid-19, a condução das atividades foi significativamente afetada. Além do semestre letivo ter sido menor do que o habitual, consequentemente acelerando o processo de desenvolvimento do trabalho, a precaução com o distanciamento social e a interrupção de aulas presenciais no campus da universidade impactou no número de voluntários. No total, foram 29 voluntários que realizaram o processo de gravação por completo, totalizando 5800 sentenças gravadas em aproximadamente 15 horas de áudio em arquivo bruto.

Figura 1: Distribuição por Sexo



Além das amostras de áudio, também coletou-se dados sobre características físicas de cada participante. Além da distribuição, também foram calculadas a variância dos dados e o desvio padrão. Os cálculos foram realizados utilizando os dados brutos e não a categorização apresentada nas figuras.

A Figura 1 demonstra de forma quantitativa a distribuição de voluntários entre os sexos. Os voluntários foram em maior parte do sexo masculino, sendo 19 (65.5%) do total, contra 10 (34.5%) voluntários do sexo feminino. características físicas.

Houve voluntários que se identificaram apenas com etnia 'Branca' e 'Pardo' das cinco etnias possíveis para seleção. A Figura 2 mostra essa distribuição. Entre essas duas etnias, majoritariamente houve a identificação com a etnia 'Branca'.

Para apresentar a distribuição de idade, criaram-se classes com intervalos de 5 anos entre 20 e 45 anos. A Figura 3 mostra essa distribuição. Observa-se que houve participação maior de pessoas com menos de 30 anos, representando aproximadamente 76% do total de voluntários (variância: 32,635 ; desvio padrão: 5,71). Em relação à apresentação da distribuição de altura criaram-se classes com intervalos de 5 centimetros de 150 até 180 centímetros ou mais. Os resultados apresentam uma variância de

Figura 2: Distribuição por Etnia

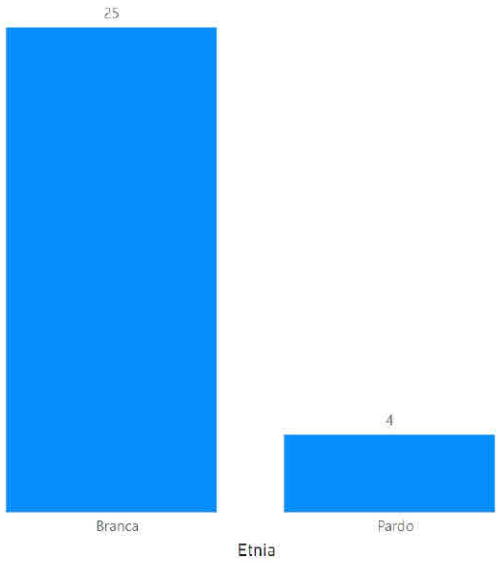
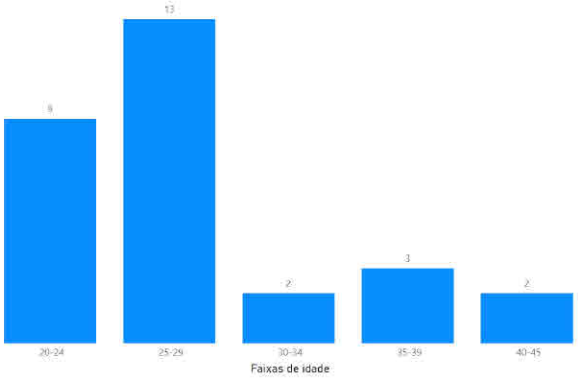


Figura 3: Distribuição por Idade



109,69 e desvio padrão de 10,47. A Figura 4 mostra essa distribuição.

Por fim, para apresentar a distribuição de peso, criaram-se classes com intervalos de 10 kilos de 40 até 99 kilos ou mais, conforme mostrado na Figura 5. Os resultados apresentam uma variância de 285,35 e desvio padrão de 16,89.

5 Discussão

Comparativamente aos trabalhos relacionados indicados, o banco de fala construído no presente trabalho pode ser considerado pequeno. Apesar dos esforços empregados pelo pesquisador para angariar o maior número possível de participantes, a pandemia do Covid-19 impactou as atividades. Ainda assim, o banco de fala possui notável valor dada a escassez de dados públicos de áudio de fala do português brasileiro com cobertura fonética abrangente. Destaca-se também, nesse aspecto, a consistência na qualidade do áudio e homogeneidade nas condições de gravação. Embora a quantidade de locutores possa ser inferior àquelas apresentadas na literatura, a consolidação do presente trabalho abre oportunidades para estudos fonéticos, de

Figura 4: Distribuição por Altura

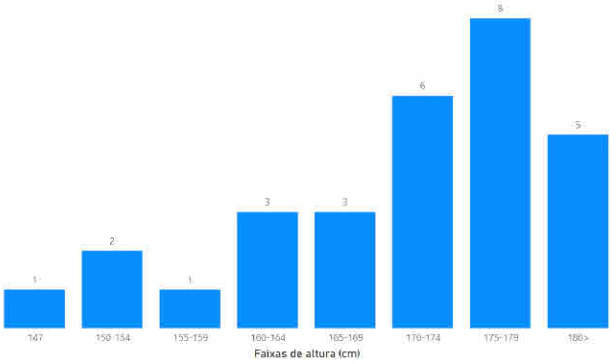
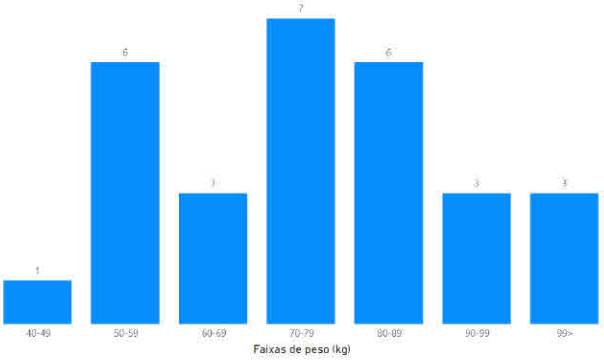


Figura 5: Distribuição por Peso



sintetização e reconhecimento de fala, dentre outros. O local de gravação possui isolamento acústico, o que foi aliado ao fato do ambiente externo (prédio e salas de aula) estarem sem utilização devido à paralização das atividades na universidade. Por fim, a escolha das frases lidas pelos voluntários, que levou como base trabalhos relacionados e a maximização da cobertura de fonemas, permite uma maior generalização de uso do banco de falas. Os dados podem ser utilizados, por exemplo, para diferentes finalidades. Uma dessas opções pode ser em relação a pesquisas que relacionem o fenótipo humano com sua voz, visto que diversos metadados foram coletados adicionalmente à voz.

Embora o banco de fala tenha sido criado, com disponibilidade pública, ainda faz-se necessário sua validação empírica. O escopo do presente trabalho foi limitado à organização e coleta de áudios, sem testes ou validações de áudio. O banco de fala deve ser testado e validado com modelos de sintetização de fala, por exemplo, para que exista comprovação de sua eficácia e qualidade. Comprovada sua eficácia, pode-se então utilizar os dados para pesquisas. Embora não exista uma validação dos dados, o banco de fala proposto pode ser ampliado a qualquer momento. Essa possibilidade de expansão pode ser conduzida como trabalhos futuros desenvolvidos na UFFS. Isso é factível dada as condições e procedimentos que são detalhados neste trabalho, bem como a existência dos equipamentos e a sala de gravação estarem à disposição para uso de acadêmicos e docentes.

Por fim, com relação aos metadados coletados dos voluntários, observou-se uma baixa variância e desvio padrão. Isso pode prejudicar trabalhos futuros que venham a explorar a relação entre identidade de voz e características físicas. Especula-se que esta baixa variância seja decorrente da proximidade social dos voluntários e da limitação geográfica onde o trabalho foi realizado. Isso ilustra a dificuldade logística de construção de bancos de fala na convocação de voluntários, na cobertura fonética e na obtenção de variabilidade entre os participantes. Essa dificuldade foi expressada também nos trabalhos relacionados por outros pesquisadores. Nesse contexto, a predominância de locutores masculinos também impacta na diversidade e validade dos dados. O mesmo efeito e viés existe em relação a aspectos sociais e geográficos, que impactaram significantemente os atributos de sexo, etnia e idade. Ainda sim, pode-se explorar a relação entre identidade de voz e características físicas com os atributos coletados, mesmo que limitados à altura e peso, por exemplo.

6 Conclusão

Esse trabalho apresentou o planejamento, execução e desenvolvimento de um banco de fala para o Português Brasileiro com foco ao suporte de estudos de sintetização de fala artificial. Além das amostras de áudio coletadas, também foram registrados dados de características físicas dos voluntários como complemento para a realização de estudos que correlacionam identidade de voz com características físicas. A variância dos dados de características físicas coletados pode ser considerada baixa, o que pode dificultar estudos dessa natureza.

Participaram do trabalho um total de 29 voluntários, resultando em 5.800 sentenças gravadas em aproximadamente 15 horas de áudio bruto. Todas as amostras foram gravadas sob as mesmas condições de gravação e acústica, o que garante homogeneidade entre as amostras, livre de ruídos e com qualidade na captura de áudio. O processo foi realizado em uma sala com isolamento acústico e equipamento de áudio profissional, o que garante um teor de qualidade para os dados da base.

Todos os dados coletados estão públicos e disponíveis para utilização para quaisquer fins educacionais, institucionais, científicos ou informativos, excluindo-se fins comerciais. Como sugestões de trabalhos futuros, pode-se repetir o experimento conduzido para coleta de audios visando-se ampliar o número de falantes, bem como a diversidade da base de fala. Trabalhos futuros podem focar, também, na validação empírica dos dados coletados, como a utilização para o treinamento de modelos sintéticos ou de aprendizado de máquina.

Referências

[1] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang. A review of deep learning based speech synthesis. 2019. URL https://www.researchgate.net/publication/336113314_A_Review_of_Deep_Learning_Based_Speech_Synthesis.

[2] Karolina Kuligowska, Paweł Kisielewicz, and Aleksandra Włodarz. Speech synthesis systems: disadvantages and limitations. 2018. URL https://www.researchgate.net/publication/325554736_Speech_synthesis_systems_Disadvantages_and_limitations.

[3] Vanessa Marquiafavel Serrani and Luis Felipe Uebel. Bancos de fala para o português brasileiro. 2011. URL https://www.linguamatica.com/index.php/linguamatica/article/view/82.

[4] Luiz Felipe Santos Vecchietti. Processamento de uma nova base de voz com aplicação em síntese de fala utilizando modelos ocultos de markov. 2015. URL http://monografias.poli.ufrj.br/monografias/monopoli10013094.pdf.

[5] Igor Macedo Quintanilha. End-to-end speech recognition applied to brazilian portuguese using deep learning. 2017. URL http://www.pee.ufrj.br/index.php/pt/producao-academica/dissertacoes-de-mestrado/2017/2016033174-end-to-end-speech-recognition-applied-to-brazilian-portuguese-using-deep-learning/file.

[6] O projeto da norma urbana linguística culta - recife (nurc/recife). https://fale.ufal.br/projeto/nurcdigital/,. Data de acesso: 21-03-2021.

[7] Miguel Oliveira Junior. Nurc digital um protocolo para a digitalização, anotação, arquivamento e disseminação do material do projeto da norma urbana linguística culta (nurc). 2016. URL https://www.researchgate.net/publication/311876616_NURC_Digital_Um_protocolo_para_a_digitalizacao_anotacao_arquivamento_e_disseminacao_do_material_do_Projeto_da_Norma_Urbana_Linguistica_Culta_NURC.

[8] O projeto da norma urbana linguística culta - rio de janeiro (nurc/rj). https://nurcrj.letras.ufrj.br/,. Data de acesso: 21-03-2021.

[9] O projeto da norma urbana linguística culta - são paulo (nurc/sp). http://nurc.fflch.usp.br/,. Data de acesso: 21-03-2021.

[10] John Garofolo, Lori Lamel, William Fisher, Jonathan Fiscus, David Pallett, Nancy Dahlgren, and Victor Zue. TIMIT Acoustic-Phonetic Continuous Speech Corpus, 1993. URL https://hdl.handle.net/11272.1/AB2/SWVENO.

[11] Abraham Alcaim. Freqüência de ocorrência dos fones e listas de frases foneticamente balanceadas no português falado no rio de janeiro. 1992. URL https://jcis.sbrt.org.br/jcis/article/view/166.

[12] Aäron van den Oord. Wavenet: A generative model for raw audio. 2016. URL https://arxiv.org/abs/1609.03499.

[13] Voice and articulation drillbook. by grant fairbanks. harper and brothers, 1941. cloth, 234 pp. *The Laryngoscope*, 51(12):1141–1141, 1941. doi: https://doi.org/10.1288/00005537-194112000-00007. URL https://onlinelibrary.wiley.com/doi/abs/10.1288/00005537-194112000-00007.

[14] Sólon do Valle. *Microfones por Sólon do Valle*. 2ª edition, 2002.

[15] Izabel Christine Seara. Estudo estatístico dos fonemas do português falado na capital de santa catarina para elaboração de frases foneticamente balanceadas. 1994. URL https://repositorio.ufsc.br/handle/123456789/112119.

[16] Jody Kreiman and Diana Sidtis. *Foundations of Voice Studies: An Interdisciplinary Approach to Voice Production and Perception*. 2011.