Pós-Graduação em Ciência da Computação

CHAINA SANTOS OLIVEIRA

Uma Arquitetura para Teste de Sistemas de Reconhecimento da Fala com Geração Automática de Áudios

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2019

CHAINA SANTOS OLIVEIRA

Uma Arquitetura para Teste de Sistemas de Reconhecimento da Fala com Geração Automática de Áudios

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Engenharia de Software

Orientador: Ricardo Bastos Cavalcante Prudêncio

Recife
2019

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

O48a Oliveira, Chaina Santos
Uma arquitetura para teste de sistemas de reconhecimento da fala com geração automática de áudios / Chaina Santos Oliveira. – 2019.
87 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2019.
Inclui referências e apêndice.

1. Engenharia de software. 2. Teste de software. 3. Sintetização da fala. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

005.1 CDD (23. ed.) UFPE- MEI 2019-071

Chaina Santos Oliveira

"Uma Arquitetura para Teste de Sistemas de Reconhecimento da Fala com Geração Automática de Áudios"

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação.

Aprovado em: 28/02/2019.

BANCA EXAMINADORA

_____

Prof. Dr. Tsang Ing Ren (Examinador Interno)
Universidade Federal de Pernambuco

_____

Prof. Dr. Márcio de Medeiros Ribeiro (Examinador Externo)
Universidade Federal de Alagoas

_____

Prof. Dr. Ricardo Bastos Cavalcante Prudêncio (Orientador)
Universidade Federal de Pernambuco

Dedico este trabalho a minha família e amigos que foram porto seguro perante as dificuldades durante este percurso.

AGRADECIMENTOS

Agradeço a todos os professores que compartilharam comigo seus conhecimentos. Sou grata especialmente ao professor Ricardo Prudêncio, pela orientação prestada, auxílio na pesquisa, disponibilidade, incentivo e apoio. Aqui lhe exprimo a minha gratidão.

A meus pais, que sempre estiveram ao meu lado e batalharam por anos para proporcionarem uma educação melhor aos seus filhos. Entendo o quanto vocês se doaram para que eu chegasse até aqui. À minha irmã Larissa, por acreditar em mim e me dar forças. Ao meu irmão Ícaro (caçula), que respeitou minha ausência e momentos em que eu não podia brincar. Família, obrigada pelo amor!

Aos demais familiares, por torcerem por mim, em especial a tia Jane e tio Donizete que são como meus segundos pais. Obrigada por sempre me apoiarem, acolherem e mostrarem, com palavras e dando exemplos, que os estudos abrem portas para um futuro melhor.

Aos amigos-irmãos que a vida acadêmica me proporcionou, em especial a Johny, Kécia, Mayra e Renan. Vocês são 10! Aos outros meus amigos, em especial à minha amiga de infância, Clécia, que está desde sempre prestando seu apoio incondicional.

A Deus, por ter colocado essas pessoas maravilhosas em meu caminho. Sem elas, eu não teria conseguido. Agradeço também por ter me dado forças e saúde!

Enfim, muito obrigada a todos que contribuíram, direta ou indiretamente, para que eu chegasse até aqui :).

RESUMO

As aplicações que utilizam sistemas de reconhecimento de fala (*speech to text* - STT) estão em ascendência nos últimos anos. Tal crescimento se deu tanto pela evolução de pesquisas acadêmicas na área, quanto pela facilidade de comunicação via fala. Esses tipos de software têm simplificado a interação entre humanos e máquinas (e.g., sistemas para *smartphones*, *smart home*, *smart cities*, etc.). Tais aplicações possuem uma variedade de usuários (nacionalidades, sotaques e gêneros diferentes) que influenciam diretamente na avaliação da qualidade de tais sistemas. Os usuários são exigentes e as diferenças anteriormente citadas devem ser levadas em consideração no momento de avaliar tais aplicações. Uma das atividades fundamentais na garantia da qualidade em aplicações que utilizam sistemas STT é o teste de SW. Para tal, faz-se necessário a utilização de técnicas que consigam reproduzir as variações da fala humana para a obtenção de resultados mais expressivos e, com isso, evitar o uso de pessoas (fala gravada) devido aos altos custos e disponibilidade. Diante disso, o uso de falas sintéticas para teste de sistemas STT seria uma opção às falas humanas devido ao seu baixo custo e praticidade de obtenção. Dado esse contexto, o presente trabalho propõe uma arquitetura para testes de sistemas STT com áudios sintetizados utilizando quatro abordagens de síntese diferentes. Para a validação do uso de áudios sintéticos como uma alternativa aos gravados, foram realizados experimentos automatizados (aplicados a sistemas de STT em *smartphones*) e baseados na opinião de pessoas (*i.e.*, teste de Turing e de qualidade). Ambos os experimentos utilizaram um ambiente real de teste de SW nas dependências do projeto CIn-Motorola.

Palavras-chaves: Sintetização da Fala. Teste de Software. Sistemas de Reconhecimento da Fala.

ABSTRACT

In recent years, applications that use speech-to-text (STT) systems are in the ascendancy. Such growth is due to the evolution of academic research in the area and to the ease of communication through speech. These softwares have simplified the interaction between humans and machines (e.g., systems for smartphones, smart home, smart cities, etc.). Such applications have a variety of users (different nationalities, accents and genres) that directly influence the quality evaluation of such systems. Users are demanding and the differences mentioned above should be taken into account when evaluating such applications. One of the fundamental activities in quality assurance in applications using STT systems is the SW test. It is necessary to use techniques that can reproduce the variations of human speech to obtain more expressive results, and thus avoid the use of people (recorded speech) due to the high costs and availability. Therefore, the use of synthetic speeches to test STT systems is an option to substitute human speech because of its low cost and practicality of obtaining. Given this context, the present work proposes an architecture for testing STT systems with audios synthesized using four different synthesis approaches. For the evaluation of the use of synthetic audios as an alternative to the recorded ones, automated experiments (applied to STT systems in smartphones) and based on the opinion of people (i.e., Turing test and quality) were made. Both experiments used a real SW test environment in the CIn-Motorola project dependencies.

Keywords: Speech Synthesis. Software Testing. Speech Recognition Systems.

LISTA DE FIGURAS

Figura 1 – Arquitetura de teste do sistema de reconhecimento de fala de Crepy, Kusnitz e Lewis (2003) 19
Figura 2 – Modelo de avaliação de um sistema de reconhecimento de fala por Crepy, Kusnitz e Lewis (2003) 19
Figura 3 – Modelo de avaliação de um sistema de reconhecimento de fala por Gandhi et al. (2004) 20
Figura 4 – Arquitetura de teste do sistema de reconhecimento de fala de Cohen et al. (2005) 21
Figura 5 – Modelo de avaliação de um sistema de reconhecimento de fala por Cohen et al. (2005) 22
Figura 6 – Sistema *text to speech* 24
Figura 7 – Modelo Fonte-Filtro de Produção da Fala (LEMMETTY, 1999) 27
Figura 8 – Modelo de Sintetizador Formante em Cascata (LEMMETTY, 1999) . . . 27
Figura 9 – Modelo de Sintetizador Formante em Paralelo (LEMMETTY, 1999) . . . 28
Figura 10 – Visualização de uma pilha WaveNet (OORD et al., 2016) 31
Figura 11 – Arquitetura Proposta 41
Figura 12 – Fluxo Geral da Arquitetura Proposta 42
Figura 13 – Pré-processamento da Lista de Sentenças 43
Figura 14 – Geração de Sentenças Equivalentes 45
Figura 15 – Filtragem de Sentenças Equivalentes 46
Figura 16 – Sintetização da Fala por Determinado Serviço 48
Figura 17 – Filtragem de Sentenças Equivalentes 49
Figura 18 – Porcentagem de sentenças corretamente transcritas 52
Figura 19 – Porcentagem de transcrições corretas e incorretas por locutor 53
Figura 20 – Porcentagem de Sucesso e Falha 54
Figura 21 – Porcentagem de transcrições corretas e incorretas, levando em consideração apenas as falhas no teste funcional 54
Figura 22 – Correlação entre as execuções dos testes funcionais com falas sintetizadas e humanas 55
Figura 23 – Histogramas dos áudios gravados 57
Figura 24 – Histogramas dos áudios sintetizados 58
Figura 25 – Correlação entre as execuções dos testes funcionais com falas sintetizadas e humanas 59
Figura 26 – Características dos voluntários que avaliaram os áudios 60
Figura 27 – Taxa de falha e sucesso geral e por tipo (humano e sintético) 61
Figura 28 – Taxa de falha e sucesso por gênero dos áudios sintéticos 62

Figura 29 – Taxa de falha e sucesso por serviço 62
Figura 30 – Taxa de falha e sucesso por serviço por gênero 63
Figura 31 – Taxa de falha e sucesso (Google Padrão e Wavenet) 64
Figura 32 – Taxa de falha e sucesso por locutor - Azure 64
Figura 33 – Taxa de falha e sucesso por locutor - Google 65
Figura 34 – Taxa de falha e sucesso por locutor - Polly 65
Figura 35 – Taxa de falha e sucesso por locutor - Watson 66
Figura 36 – Taxa de falha e sucesso por locutor - Google Padrão e Wavenet 66
Figura 37 – Taxa de falha e sucesso por linguagem de todos os áudios sintéticos . . 67
Figura 38 – Fluência dos voluntários x Respostas corretas no teste de Turing 67
Figura 39 – Qualidade geral (humano e sintetizado) 68
Figura 40 – Qualidade por gênero 68
Figura 41 – Qualidade por serviço 69
Figura 42 – Sucesso (Turing) x Qualidade 69
Figura 43 – Taxa de falha e sucesso geral e por execução 71
Figura 44 – Taxa de falha e sucesso por gênero geral (nas cinco execuções) 72
Figura 45 – Taxa de falha e sucesso por serviço 72
Figura 46 – Taxa de falha e sucesso por serviço por gênero 73
Figura 47 – *Word Error Rate* (WER) por sentença 73
Figura 48 – Correlação das WERs dos áudios humanos e sintéticos 74
Figura 49 – Qualidade x Sucesso Automático (transcrição) 75
Figura 50 – Taxa de falha e sucesso geral e por execução (funcional) 75
Figura 51 – Taxa de falha e sucesso por gênero geral (nas cinco execuções) 76
Figura 52 – Taxa de falha e sucesso por serviço 76
Figura 53 – Taxa de falha e sucesso por serviço por gênero 77

LISTA DE TABELAS

Tabela 1 – Vozes em Inglês Disponíveis no Amazon Polly 32
Tabela 2 – Exemplos de Entrada para o Polly 33
Tabela 3 – Vozes em Inglês Disponíveis no Google *Text to Speech* 34
Tabela 4 – Exemplos de Entrada para o Google TTS 35
Tabela 5 – Vozes em Inglês Disponíveis na IBM Watson 36
Tabela 6 – Exemplos de Entrada para o IBM Watson 36
Tabela 7 – Vozes em Inglês Disponíveis no Microsoft Azure 37
Tabela 8 – Exemplos de Entrada para o Microsoft Azure TTS 37
Tabela 9 – Geração de Sentenças Equivalentes 46
Tabela 10 – Filtragem de Sentenças 47
Tabela 11 – Quantidade de áudios humanos (gravados) e sintetizados utilizados . . 57
Tabela 12 – Significado das labels de falha e sucesso 61
Tabela 13 – Vozes disponíveis no Amazon Polly 83
Tabela 14 – Vozes disponíveis no Google *Text-to-Speech* 84
Tabela 15 – Vozes disponíveis na IBM Watson 85
Tabela 16 – Vozes disponíveis no Microsoft Azure Bing Speech 86

LISTA DE ABREVIATURAS E SIGLAS

HMM	*Hidden Markov Model*
NLTK	*Natural Language Toolkit*
SSML	*Speech Synthesis Markup Language*
STT	*Speech to Text*
TTS	*Text to Speech*
WER	*Word Error Rate*

SUMÁRIO

1 INTRODUÇÃO 14
1.1 PROBLEMA E MOTIVAÇÃO 14
1.2 OBJETIVOS 15
1.3 TRABALHO REALIZADO 15
1.4 ORGANIZAÇÃO DO TRABALHO 15

2 TESTES EM SISTEMAS DE RECONHECIMENTO DE FALA . . . 17
2.1 SISTEMAS SISTEMAS DE RECONHECIMENTO DE FALA 17
2.2 ARQUITETURAS DE TESTE 18
2.3 CASOS DE TESTE 22
2.3.1 Áudios Gravados 23
2.3.2 Áudios Extraídos 23
2.3.3 Áudios Sintetizados 24
2.4 CONSIDERAÇÕES FINAIS 24

3 SINTETIZAÇÃO DE FALA 26
3.1 MÉTODOS DE SINTETIZAÇÃO *TEXT TO SPEECH* 26
3.1.1 *Formant* 26
3.1.2 *Articulatory* 28
3.1.3 *Concatenative* 29
3.1.4 Paramétrico Estatístico 30
3.1.5 Síntese Híbrida 30
3.1.6 WaveNet 30
3.2 SERVIÇOS DE SINTETIZAÇÃO 31
3.2.1 Amazon Polly 31
3.2.2 Google *Text to Speech* 33
3.2.3 IBM Watson *Text to Speech* 35
3.2.4 Microsoft Azure *Text to Speech* 36
3.3 CONSIDERAÇÕES FINAIS 38

4 ARQUITETURA DE TESTE DE SISTEMAS *SPEECH TO TEXT* . 39
4.1 CONTEXTO GERAL E OBJETIVO 39
4.2 ARQUITETURA PROPOSTA 41
4.3 IMPLEMENTAÇÃO 43
4.3.1 Pré-processamento das Sentenças 43
4.3.2 Geração de Sentenças Equivalentes 44

4.3.3 Filtragem de Sentenças . . . 46
4.3.4 Sintetização da Fala . . . 47
4.3.5 Teste de Sistema *Speech to Text* . . . 48
4.4 CONSIDERAÇÕES FINAIS . . . 50

5 EXPERIMENTOS E RESULTADOS . . . 51
5.1 EXPERIMENTO I . . . 51
5.1.1 Protocolo Experimental . . . 51
5.1.2 Resultados . . . 52
5.2 EXPERIMENTO II . . . 55
5.2.1 Avaliação Subjetiva . . . 56
5.2.1.1 Protocolo Experimental . . . 56
5.2.1.2 Resultados . . . 60
5.2.2 Avaliação Objetiva . . . 70
5.2.2.1 Protocolo Experimental . . . 70
5.2.2.2 Resultados . . . 70
5.3 CONSIDERAÇÕES FINAIS . . . 77

6 CONCLUSÕES E TRABALHOS FUTUROS . . . 78
6.1 CONTRIBUIÇÕES . . . 78
6.2 TRABALHOS FUTUROS . . . 78

REFERÊNCIAS . . . 80

APÊNDICE A – VOZES DISPONÍVEIS NOS SERVIÇOS DE SÍNTESE . . . 83

1 INTRODUÇÃO

Neste capítulo introdutório será mostrado, em linhas gerais, o contexto ao qual o presente trabalho está inserido e um resumo de suas contribuições obtidas com a sua elaboração.

1.1 PROBLEMA E MOTIVAÇÃO

A comunicação via fala é uma das mais usadas no mundo atual. Não é por acaso que o meio tecnológico tem investido em aplicações que a utiliza como forma de interação com o usuário. Isso, bem como a evolução em pesquisas acadêmicas relacionadas ao reconhecimento da fala são alguns dos fatores responsáveis pelo crescente desenvolvimento dessas aplicações (YU; DENG, 2015).

Por outro lado, como qualquer outro software, o que o usuário deseja é que ele funcione bem, que atenda às suas requisições e retorne o resultado esperado, conforme as especificações. No caso de um sistema de reconhecimento de fala (*Speech to Text* (STT)), é esperado que, no mínimo, ele seja capaz de entender o que foi falado. Dentro desse contexto, os responsáveis por desenvolver este tipo de sistema lidam com uma série de desafios, uma vez que ele pode ser usado por pessoas de sotaques, gêneros e nacionalidades diferentes.

Diante disso, para que o produto (com um sistema STT) chegue às mãos do usuário final com boa qualidade, neste caso, com bom reconhecimento da fala, é necessário que ele seja testado com a maior quantidade de variações de falas possível. Tais testes podem ser feitos de forma manual ou automática. Se forem feitos manualmente, são necessárias várias pessoas, com as variações citadas acima para fazer as execuções (dos testes). Esta opção é praticamente inviável, uma vez que demanda custo de tempo e financeiro alto.

A outra alternativa são os testes automatizados, onde já não são necessárias as pessoas com tais especificidades todo momento que for fazer os testes. No entanto, são necessárias falas humanas de forma gravada, com os comandos a serem testados. O lado positivo é que as falas gravadas podem ser reutilizadas. O negativo é em relação à aquisição delas, que, novamente, é custoso em questão de tempo e financeira. Tal custo é decorrente, da contratação de terceiros para gravar as frases específicas, como também dos equipamentos, ambiente e tempo necessários para a gravação.

Há, ainda, o desenvolvimento de pesquisas e serviços de sintetização da fala que vem crescendo. Tais serviços oferecem alguns tipos de variações, como nacionalidades, gêneros e sotaques diferentes. Desta forma, poderia-se optar por utilizar falas sintetizadas às gravadas nos testes de sistemas de reconhecimento de fala.

1.2 OBJETIVOS

O objetivo principal deste trabalho é desenvolver uma arquitetura de teste de sistema de reconhecimento de fala utilizando áudios sintetizados. Para alcançar tal objetivo, alguns específicos foram realizados:

- Implementar a arquitetura para teste de sistemas de reconhecimento de fala usando áudios sintetizados;
- Mostrar que os áudios gravados podem ser substituídos pelos sintéticos nos testes de sistemas STT;
- Realizar experimentos a fim de validar o trabalho desenvolvido;

1.3 TRABALHO REALIZADO

Este trabalho propõe uma arquitetura para teste de sistema de reconhecimento de fala, utilizando áudios sintéticos. Ela mostra que os áudios gravados podem ser substituídos pelos sintéticos neste tipo de teste. Para tal, foram realizadas avaliações tanto objetivas, quanto subjetivas.

As avaliações objetivas foram baseadas na opinião dos usuários sobre a naturalidade e qualidade dos áudios sintetizados (*i.e.*, teste de Turing e de qualidade). Já as subjetivas foram baseadas na execução de testes automáticos num *smartphone* nas dependências do projeto CIn-Motorola. Essas últimas visavam verificar a acurácia de transcrição dos áudios sintetizados e sua correlação com os resultados dos gravados.

1.4 ORGANIZAÇÃO DO TRABALHO

Este trabalho está organizado em seis capítulos:

- Capítulo 1: capítulo introdutório.
- Capítulo 2: alguns trabalhos similares ao proposto foram apresentados, bem como os possíveis tipos de áudios que podem ser usados nos testes de sistemas de reconhecimento da fala.
- Capítulo 3: foi abordado sobre os métodos de sintetização existentes, bem como quatro dos serviços disponíveis no mercado.
- Capítulo 4: todos os módulos da arquitetura proposta (o pré-processamento, sintetização, geração de sentenças equivalentes, filtragem de sentenças e teste de sistemas de reconhecimento da fala), e as respectivas implementações foram detalhadas.
- Capítulo 5: os experimentos, protocolos e resultados foram detalhados.

- Capítulo 6: a conclusão do presente trabalho e sugestões para pesquisas futuras foram feitas.

2 TESTES EM SISTEMAS DE RECONHECIMENTO DE FALA

A fala é uma das formas de comunicação mais usadas e eficientes (VIMALA; RADHA, 2012). Em virtude disso, o meio tecnológico tem investido cada vez mais em tipos de iterações entre os usuários e aplicações por meio de voz (YANG; HONG, 2015). Isso tem ganhado popularidade e por outro lado, levantado uma série de desafios tanto no meio acadêmico, quanto de mercado.

Um dos maiores desafios é a construção de sistemas de reconhecimento de fala inteligentes, que além de serem capazes de entender aquilo que o usuário diz, dão o retorno correto, o resultado esperado (ERRATTAHI; HANNANI; OUAHMANE, 2015). Desta forma, no processo de desenvolvimento desse tipo de sistema, é necessário que sejam feitos testes robustos, explorando os mais variados tipos de cenário e com várias vozes diferentes. Neste capítulo, exploraremos os sistemas STT, testes para esse tipo de sistema, e casos de teste utilizados nesse processo.

2.1 SISTEMAS SISTEMAS DE RECONHECIMENTO DE FALA

Um sistema de reconhecimento da fala é capaz de entender o que alguém falou e traduzir para um formato de leitura de máquina. Normalmente essa tradução é feita da fala para texto, o que é chamando de *speech to text* (VIMALA; RADHA, 2012; MANASWI, 2018). Esse tipo de comunicação entre o usuário e o sistema evita que outros métodos de comunicação sejam usados, como o teclado, botões, tela, ou até mesmo a comunicação via gestos (CREPY; KUSNITZ; LEWIS, 2003).

Dessa forma, tecnologias que permitem a comunicação através da fala vêm mudando a forma que vivemos, uma vez que, em alguns casos essa é primeira forma de interação entre os humanos e máquinas. O crescimento no desenvolvimento de tecnologias assim tem acontecido devido a evolução do poder de processamento computacional, permitindo o treinamento de modelos de reconhecimento de fala cada vez mais robustos. Deve-se também ao grande volume de dados disponíveis para o treinamento de tais modelos. E ainda, à popularidade de dispositivos móveis, versáteis, de sistemas de infotenimento[1] em veículos e outros, onde a comunicação via teclado não é tão conveniente (YU; DENG, 2015).

O crescimento no uso de tais aplicações também é reflexo da melhoria e aperfeiçoamento da comunicação entre o humano e a máquina. No carro que tem um sistema assim, por exemplo, o usuário pode pedir pra tocar música, perguntar alguma informação sem que seja necessário tanto esforço, ele precisa apenas falar. Nos jogos como Xbox, este tipo de comunicação também é permitida. Falando em dispositivos como *smartphones*, já é possível fazer uma série de tarefas conversando com aparelho: pesquisas por informações

[1] É uma aglutinação dos termos informação e entretenimento. Vem do inglês *infotainment*.

na internet, por restaurantes, abrir aplicativos, fazer algumas configurações, dentre outras. Existem, ainda, os assistentes virtuais, que se tornaram mais populares depois que a Apple lançou a Siri (MANASWI, 2018).

Manaswi (2018) também cita algumas vantagens de usar sistemas *speech to text*, dentre elas, a acessibilidade para pessoas cegas, que podem controlar dispositivos diferentes apenas usando a voz. Outra é a possibilidade de manter registros de uma reunião em forma de texto, sem ser necessário digitar, apenas fazendo a transcrição do que foi falado. Também tem a possibilidade de fazer tradução de palavras/frases para outra língua, facilitando a comunicação com outras pessoas de outras nacionalidades.

Assim, pode-se observar que o uso de sistemas de comunicação através da voz vem se consolidando cada vez mais no mercado. A cada dia, mais aplicações têm surgido ou se adequando a esta nova realidade. Paralelamente, surge a necessidade de prover sistemas de qualidade, de forma que continuem a conquistar mais usuários e os mantenha satisfeitos com o uso. Nesse contexto, surge a importância de garantir o bom funcionamento do software realizando testes cada vez mais robustos e com alta cobertura.

2.2 ARQUITETURAS DE TESTE

Um dos maiores desafios para quem desenvolve sistemas de reconhecimento de fala é garantir a qualidade. Tal qualidade se refere à capacidade que o software tem de reconhecer corretamente aquilo que foi falado. Alguns, ainda, são interligados com outras aplicações para, além de entender o que foi falado, dar algum retorno para certa solicitação do usuário (e.g: sistemas em celulares). Nesse caso, assim que algo é falado, antes que o sistema retorne a resposta, é necessário que ele entenda o que foi dito. Dessa forma, são duas etapas principais que devem ser analisadas nos testes desse tipo de sistema: a transcrição da fala pra texto e a resposta dada para a solicitação feita.

Na primeira etapa, assim que algo é falado, o sistema converte o áudio para uma linguagem que a máquina entenda (*i.e.*: texto). Na segunda, o software interpreta o que foi convertido e deverá dar alguma resposta. Um bug de software pode ocorrer nos seguintes cenários: (1) se a conversão foi incorreta, pois houve um erro de entendimento do que foi falado; (2) se a conversão foi correta, mas a resposta final dada não é a esperada, pois o sistema interpretou de forma errada o que foi solicitado. Já a situação de acerto ocorre quando a conversão foi correta e a resposta dada é igual à esperada.

A indústria tem desenvolvido uma série de trabalhos voltados à avaliação de sistemas de reconhecimento de fala. Crepy, Kusnitz e Lewis (2003) desenvolveram uma arquitetura, onde, um áudio é sintetizado por um software e um sistema *speech to text* é testado. Na Figura 1, é possível ver um cenário da solução proposta. Existem dois computadores, um com o software de sintetização (1) e com uma caixa de som; e outro (2) com o microfone e o software de reconhecimento de fala. O primeiro, além de sintetizar os áudios, vai tocá-los, de forma que o segundo possa ouvir e realizar a tarefa de conversão do áudio para

texto e fazer as análises. Por fim, um relatório referente a tais análises é gerado. Os passos de execução são ilustrados no fluxograma da Figura 2.

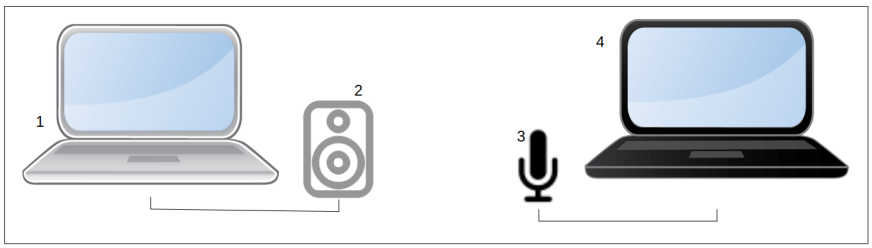


Figura 1 – Arquitetura de teste do sistema de reconhecimento de fala de Crepy, Kusnitz e Lewis (2003)

O relatório contém uma série de informações como o número de inserções (i.e., palavras que não estão em A e estão em B.), deleções (i.e., palavras que estão em A e não estão em B.) e substituições (i.e, palavras que foram entendidas de forma errada, elas estão em A e teoricamente em B, mas escritas de formas diferentes.). Por fim, a taxa de erro de cada palavra é calculada. Tal taxa é chamada de *Word Error Rate* (WER) no inglês, Equação 2.1 (CREPY; KUSNITZ; LEWIS, 2003).

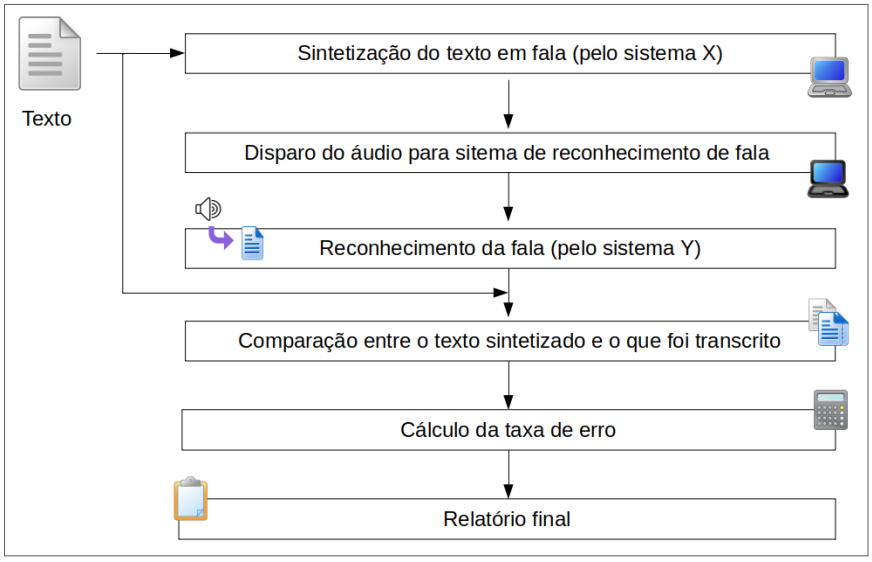


Figura 2 – Modelo de avaliação de um sistema de reconhecimento de fala por Crepy, Kusnitz e Lewis (2003)

$$W = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C} \tag{2.1}$$

Onde:

- S = número de substituições
- D = número de deleções
- I = número de inserções

- C = número de palavras corretas

Outro método de avaliação de um sistema de reconhecimento de fala foi implementando por Gandhi et al. (2004). Nele, diferentemente do descrito acima, o que se analisa é a acurácia de forma geral através de *logs* e de segmentos de áudios armazenados. O sistema de reconhecimento a ser avaliado pode estar em um computador ou outro em outro dispositivo, desde que seja capaz de gerar *logs* que têm, entre as informações, aquilo foi falado por alguém e entendido pelo sistema, bem como data e hora de certa entrada.

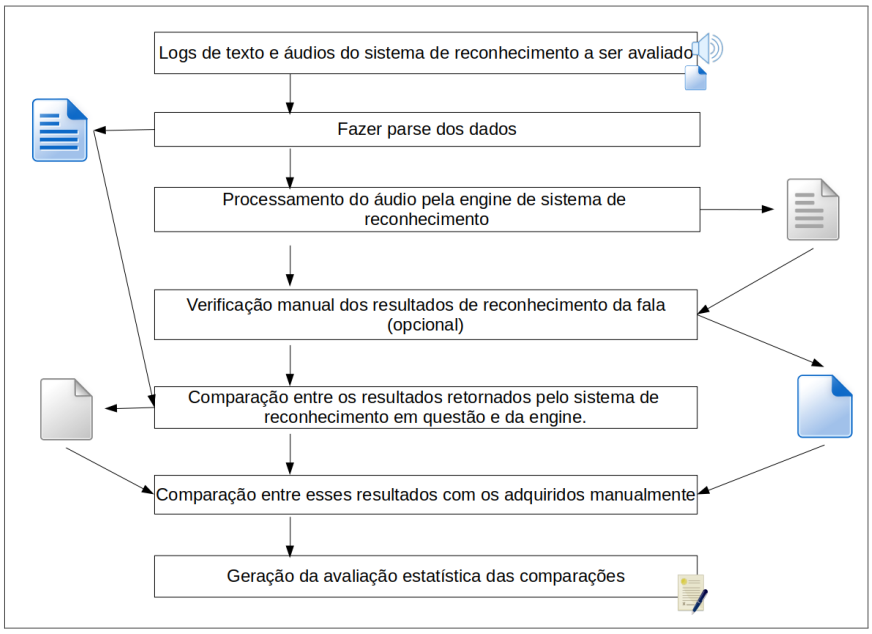


Figura 3 – Modelo de avaliação de um sistema de reconhecimento de fala por Gandhi et al. (2004)

A Figura 3 ilustra como se dá o processo de avaliação, desenvolvido por Gandhi et al. (2004), de um sistema de reconhecimento de fala. No primeiro momento, se tem como entrada os *logs* de áudio e de texto do sistema a ser avaliado. Estes últimos são passados por um parser com a finalidade de somente extrair as informações necessárias como a transcrição, a data e hora da entrada. Depois, os áudios armazenados são passados por um sistema de reconhecimento de fala referência, chamado de engine. Posteriormente, tais resultados de reconhecimento podem ser analisados manualmente. Em seguida, é feita uma comparação entre os resultados adquiridos pela transcrição do áudio feita pelo sistema referência e a armazenada no *log* do sistema de reconhecimento avaliado. Aquilo que resulta nessa etapa pode ser comparado com o que foi feito manualmente. Assim, é feita uma avaliação estatística das comparações.

Algumas das questões avaliadas são as taxas de reconhecimentos corretos e incorretos de forma geral. As taxas de tais reconhecimentos, não levando em consideração os erros causados por fatores externos. E, também é avaliado o número de reconhecimentos corretos de determinadas palavras.

Cohen et al. (2005) desenvolveram o método de teste automático de um sistema de reconhecimento de fala em um celular. Os áudios utilizados nos testes eram gravados por diversas pessoas de vários países, com sotaques distintos e em ambientes diferentes (i.e, sala silenciosa, ambiente ruidoso). O objetivo era testar vários tipos de entrada, simulando o ambiente real.

Na Figura 4, é possível ver a arquitetura de teste de tal sistema. Nela, há um computador (1) com duas caixas de som (2 e 3), uma poderá ser usada pra reproduzir os áudios gravados que serão testados no sistema, e outra poderá reproduzir ruídos, simulando um ambiente não silencioso, dependendo do tipo de teste. Também tem um celular (4) contactado ao computador, com dois tipos de conexões. Uma delas (5) indica que é possível fazer transferência direta das ondas sonoras do computador diretamente para o celular, sem ser necessário utilizar a caixa de som. A outra (6) indica a possível transferência de dados do celular para a máquina, bem como de alguma requisição feita pelo aparelho móvel.

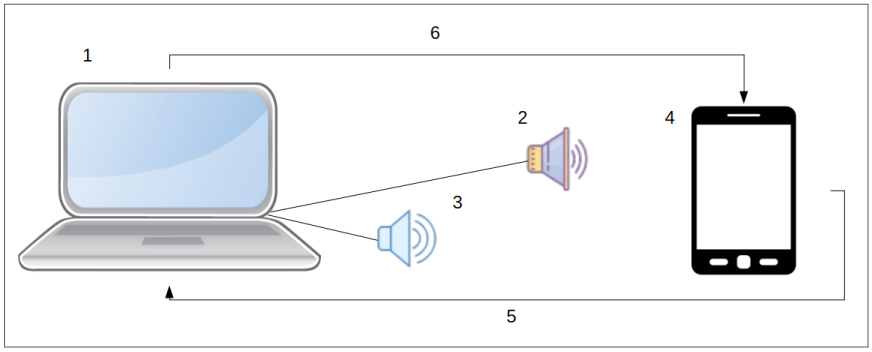


Figura 4 – Arquitetura de teste do sistema de reconhecimento de fala de Cohen et al. (2005)

O processo de teste proposto se dá da seguinte maneira: uma voz em forma de áudio é reproduzida para o celular, a resposta dada pelo aparelho é coletada e posteriormente armazenada junto com a entrada. Depois, é verificado se durante o teste é necessário o disparo de mais alguma fala. Se for, os passos descritos anteriormente são repetidos. Se não for, os resultados do teste são dado (Figura 5). É verificado se o que foi retornado está de acordo com o esperado para determinada entrada, assim é calculada porcentagem de falha e sucesso do sistema. Outro aspecto avaliado é a velocidade de resposta do sistema.

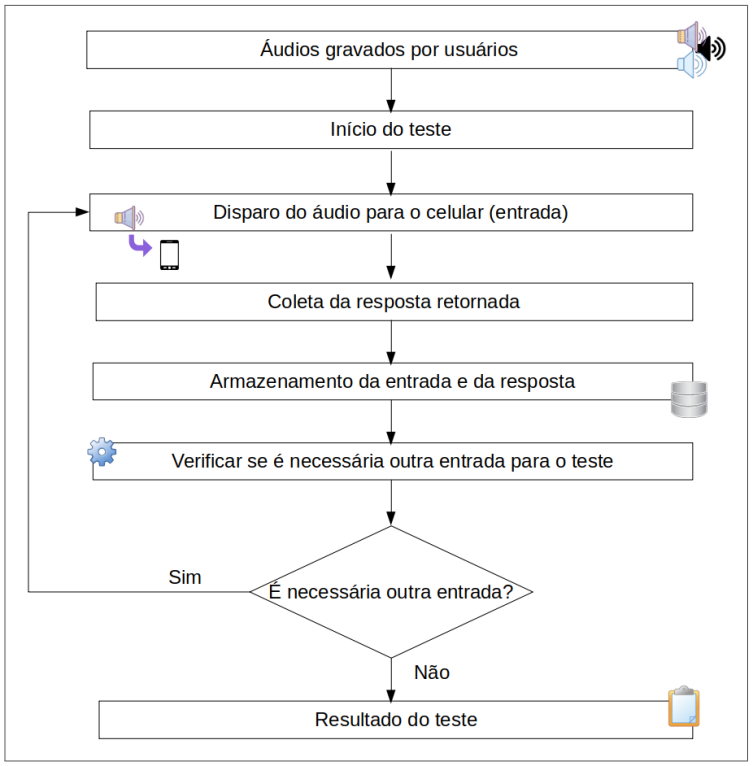


Figura 5 – Modelo de avaliação de um sistema de reconhecimento de fala por Cohen et al. (2005)

Cada uma das arquiteturas apresentadas tem suas particularidades. A primeira, desenvolvida por Crepy, Kusnitz e Lewis (2003), testa um sistema de reconhecimento de fala em um computador e no final, é gerado um relatório de acordo com o que foi transcrito corretamente, incorretamente, não transcrito; e também a WER é calculada. Os áudios utilizados em tal arquitetura foram sintetizados por outro software, diferentemente da criada por Cohen et al. (2005) que utilizou áudios humanos para testar um sistema STT de um celular. O relatório final continha a taxa de falha e o tempo de resposta dado pelo sistema.

Já o trabalho desenvolvido por Gandhi et al. (2004), analisava o reconhecimento da fala de sistema através dos *logs* de texto com as transcrições e dos áudios que foram transcritos, tendo como base um sistema referência que era utilizado no momento da análise. Os resultados dos *logs* eram comparados com os transcritos pelo sistema referência e também poderia ser feita uma verificação manual. Ao final, também era gerado um relatório com as transcrições corretas, incorretas, levando, ou não, em consideração fatores externos.

2.3 CASOS DE TESTE

Testar um sistema de reconhecimento de fala automaticamente é mais eficiente do que manualmente, uma vez que não é necessário a mesma quantidade de pessoas para fazer a execução da mesma quantidade testes. No entanto, ao contrário de outros tipos de

testes, para fazer a execução em um sistema assim, de forma automática, é necessário ter alguns áudios pré-gerados. Existem diferentes maneiras conseguir estes áudios, gravando, sintetizando (CREPY; KUSNITZ; LEWIS, 2003; COHEN et al., 2005) e extraindo, por exemplo. Essas diferentes formas de aquisição de áudios serão exploradas nas subseções seguintes.

2.3.1 Áudios Gravados

É uma das maneiras mais comuns de gerar áudios. Uma pessoa grava a frase que precisa ser testada e salva. Geralmente, isso deve ser feito em um ambiente sem ruído, com um microfone apropriado para evitar gravar áudios ruidosos, a menos que sejam áudios desse tipo que ser deseja (com ruído). Depois disso, dependendo do objetivo, eles precisam ser selecionados e pós-processados. Nesses passos, aqueles com baixa qualidade podem ser descartados, e outros terem o volume e nível de ruído ajustados, por exemplo.

Mesmo sendo uma maneira relativamente simples de obter esses áudios, é custosa porque tem muitos passos e requer que as pessoas os gravem. Quando é necessário testar vozes de pessoas de lugares distintos, com diferentes sotaques e gênero, torna-se pior porque é mais trabalhoso conseguir um número considerável de pessoas com tais diferenças.

2.3.2 Áudios Extraídos

Uma alternativa para os áudios gravados é a extração de determinadas partes de um vídeo. Isso pode ser feito da seguinte forma: dado um texto, o sistema de extração procura por ele em vídeos no YouTube[2], baixa os sons relacionados àqueles vídeos e corta o trecho correspondente à entrada fornecida. Por exemplo: dando a frase "qual a previsão do tempo?", o sistema busca por vídeos com essa sentença, faz o download de seus áudios e os corta, retornando apenas o trecho buscado, no caso: "qual a previsão do tempo?".

Essa busca no YouTube não é tão precisa. Às vezes, os vídeos retornados só contêm parte da frase, ou simplesmente não a tem. Uma opção é utilizar uma plataforma secundária que busca por vídeos do YouTube com mais precisão como o YouGlish[3]. Todos os vídeos retornados por essa plataforma realmente têm a frase buscada, ele filtra a busca de forma mais eficiente. As etapas que vêm depois da busca são iguais às descritas no parágrafo anterior, é feito o download do áudio e corte da frase correspondente.

Extrair áudios é mais eficiente e menos custoso, financeiramente falando, do que a opção descrita na Seção 2.3.1, porque não são necessárias pessoas para gravar. Em contrapartida, nem todos os textos são encontrados em vídeos. Frases complexas e longas, às vezes, podem não ser localizadas. Desta forma, a extração tem baixa cobertura. Além disso, os áudios retornados podem ser muito ruidosos. A frase pode ser encontrada em um vídeo gravado em uma festa, por exemplo. Então, assim como a obtenção de fala através da gravação, neste processo (de extração) também é necessário fazer pós-processamento.

[2] https://www.youtube.com
[3] https://youglish.com

2.3.3 Áudios Sintetizados

Pesquisadores vêm trabalhando no processo de aquisição de fala (dado determinado texto), que seria uma alternativa às gravadas e extraídas. Uma opção para executar esse processo é sintetizando as mesmas (ZEN; TOKUDA; BLACK, 2009).

Sintetizar a fala (Figura 6) diz respeito à sua aquisição utilizando uma forma artificial, sem que seja necessário uma pessoa falando naquele momento. Uma maneira de fazer isso é fornecendo um texto, no caso, uma sentença, e o software responsável por fazer a síntese, chamado sintetizador, se encarregará de dar uma saída em forma de áudio. Este tipo de sistema também pode ser chamado de *Text to Speech* (TTS) (SIDDHI; VERGHESE; BHAVIK, 2017).

Figura 6 – Sistema *text to speech*

Existem alguns métodos de sintetização de áudios: *Formant, Articulatory* e *Concatenative* (SIDDHI; VERGHESE; BHAVIK, 2017) e outros baseadas em redes neurais profundas, WaveNet, (OORD et al., 2016). A comparação entre eles, bem como vantagens e desvantagens, serão exploradas no Capítulo 3.

2.4 CONSIDERAÇÕES FINAIS

Neste capítulo foi apresentado sistema de reconhecimento de fala, bem como algumas arquiteturas desenvolvidas para testar tal tipo de software. Este tipo de sistema pode estar em um computador, bem como em outro tipo de aparelho, como celular. Algumas das formas de testar foca apenas em analisar a transcrição da fala em texto. Já outras, também avalia a resposta dada pelo sistema.

Além disso, foram apresentadas formas de aquisição de áudio para testar sistemas STT. Uma delas é através da gravação por um humano que, apesar de comum, é uma alternativa custosa quando se almeja uma variedade de falas de pessoas de diferentes gêneros, sotaques e nacionalidades. Outra é a extração em vídeos que, apesar de não ser tão custosa, tem baixa cobertura, ou seja, nem sempre todas as frases procuradas são encontradas e muito menos, com diferentes locutores. Assim, a aquisição de áudios através de sintetização vem ganhado espaço. Além de ser possível de adquirir falas com diversos

tipos de configurações, assegurando uma cobertura considerável, a sintetização não é uma alternativa custosa. No capítulo 3 serão apresentadas métodos e serviços de sintetização da fala.

3 SINTETIZAÇÃO DE FALA

O processo de sintetização vai além de produzir fala através de texto. É necessário que as falas adquiridas tenham boa qualidade. Essa característica, segundo Tabet e Boughazi (2011), pode ser medida analisando a inteligibilidade e a naturalidade da saída. Sendo que inteligibilidade diz respeito ao quão entendível a fala é; e naturalidade, ao quão parecida da voz humana, a sintetizada é.

Desta forma, existem métodos que fazem a sintetização TTS, visando atingir àqueles requisitos e/ou melhorar, de alguma forma, os previamente existentes. Eles são divididos nos seguintes grupos: *Formant, Articulatory, Concatenative*, Paramétrico Estatístico e Híbridos (SIDDHI; VERGHESE; BHAVIK, 2017; TABET; BOUGHAZI, 2011). Existe, ainda, uma técnica mais recente, baseada em redes neurais profundas, que é chamada *WaveNet* (OORD et al., 2016).

Algumas empresas também têm trabalhado na oferta de serviços de sintetização de voz. Uma delas é a Google, que deixa claro que algumas das falas produzidas por seu sistema são oriundas de processamento envolvendo WaveNet. Além dessa empresa, a Amazon, IBM e Microsoft estão desenvolvendo serviços no mesmo ramo, mas não deixam claro sobre o método de sintetização utilizado.

3.1 MÉTODOS DE SINTETIZAÇÃO *TEXT TO SPEECH*

A sintetização da fala pode ser oriunda de vários métodos. Todos possuem prós e contras e o uso de qualquer um vai depender dos requisitos da aplicação, bem como dos recursos disponíveis para o desenvolvimento e uso. Nas subseções seguintes, é possível ver com mais detalhes sobre cada um deles, bem como suas vantagens e desvantagens.

3.1.1 *Formant*

O método de síntese de voz de formantes, do inglês *formant*, é baseado em um modelo de produção da fala chamado Fonte-Filtro. Tal modelo é constituído por uma fonte sonora, que gera um sinal de voz ou ruído, e por um filtro acústico, que simula o trato vocal humano. Assim, conforme é mostrado na Figura 7, o sinal sonoro será filtrado com o filtro, que contém ressonantes, e ele se responsabilizará por modelar e produzir o sinal da fala, de acordo com alguns parâmetros passados. As ressonâncias desse filtro são chamadas formantes (LEMMETTY, 1999).

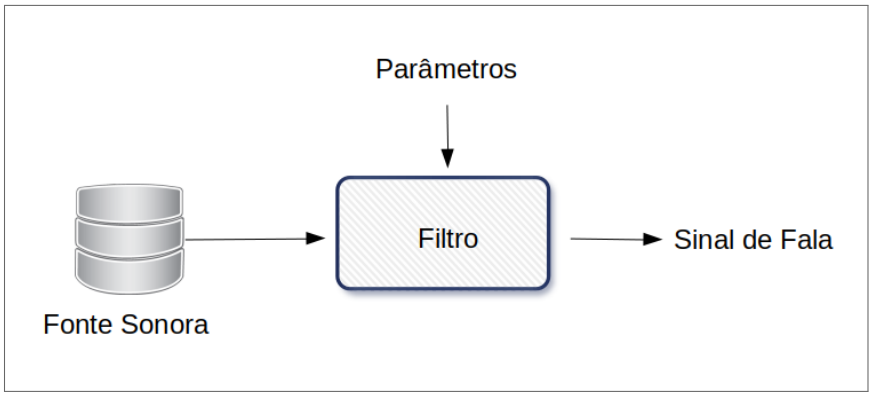


Figura 7 – Modelo Fonte-Filtro de Produção da Fala (LEMMETTY, 1999)

Como esta abordagem de sintetização é baseada em um modelo de produção da fala, não é utilizada uma amostra de vozes para fazer tal trabalho. Ao invés disso, algumas regras são escritas por linguistas para que alguns parâmetros necessários para a síntese sejam gerados (TABET; BOUGHAZI, 2011). Há duas estruturas básicas usadas na síntese de formantes, cascata e paralela. Também existem os métodos que utilizam as duas para obter melhor performance (LUKOSE; UPADHYA, 2017).

Sintetizador de Formantes em Cascata

O sintetizador em cascata é constituído por um conjunto de ressonadores em série (Figura 8), onde a saída de um é a entrada de outro e os parâmetros de entrada são apenas as frequências de formantes. Tal estrutura é simples de implementar e tem melhor desempenho em sons não nasais, devido a necessidade de menos parâmetros de controle. Por outro lado, não tem bom desempenho na modelagem de consoantes chamadas fricativas[1] (LEMMETTY, 1999).

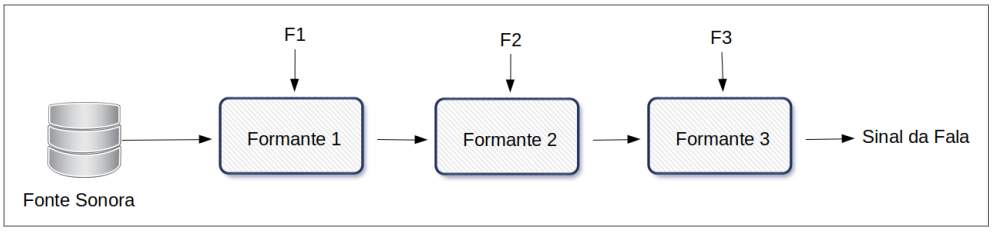


Figura 8 – Modelo de Sintetizador Formante em Cascata (LEMMETTY, 1999)

Sintetizador de Formantes em Paralelo

É formado por ressonadores em paralelo, onde as saídas de todos são somadas. Além das frequências de formantes, este tipo de sintetizador requer a largura de banda. É permitido que seja adicionado algum ressonador para a produção de sons nasais. Assim,

[1] Consoantes que são produzidas através da passagem de ar por um canal estreito.

tal estrutura tem melhor desempenho para a produção desse tipo de som. É também melhor que o cascata para a produção de consoantes fricativas e oclusivas[2]. No entanto, algumas vogais não podem ser modeladas quando este tipo de sintetizador é o utilizado (LEMMETTY, 1999).

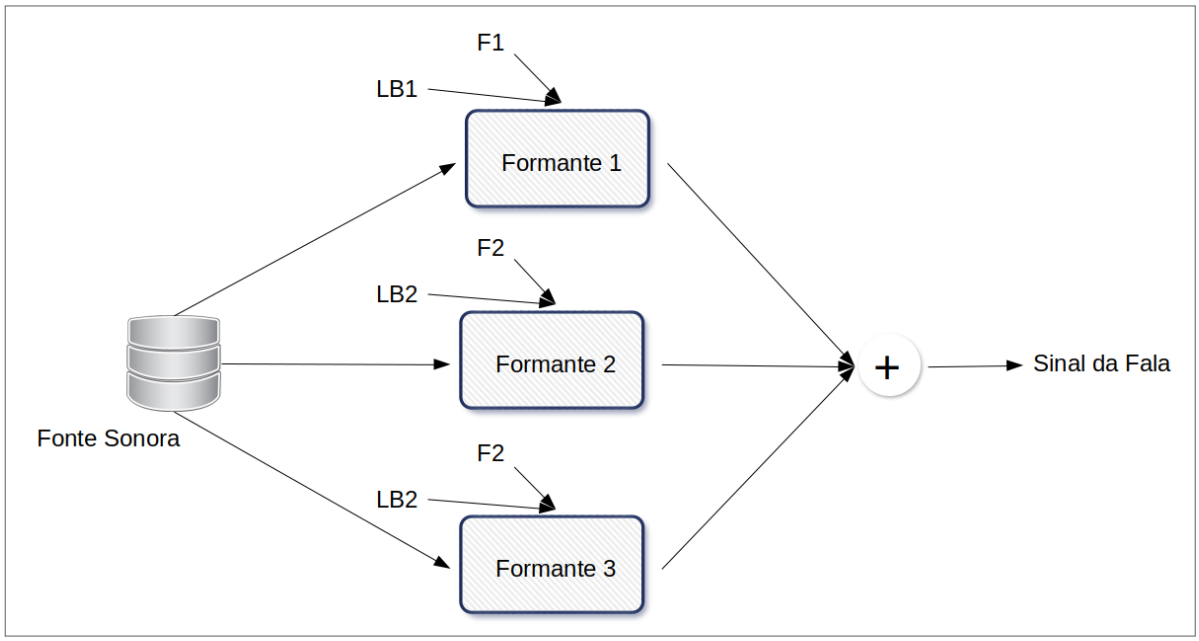


Figura 9 – Modelo de Sintetizador Formante em Paralelo (LEMMETTY, 1999)

Outro aspecto da sintetização de formantes é que ela fornece vários áudios, o que a torna mais flexível que outros tipos, que a *Concatenative*, por exemplo (LEMMETTY, 1999). Ela também tem como característica gerar áudios meio robóticos. Isso faz com o que o objetivo de naturalidade não seja atingido. Por outro lado, a fala sintetizada é bem compreensível e o custo de memória é consideravelmente baixo, uma vez que, conforme dito antes, não é necessária uma amostra de sons para treinamento (TABET; BOUGHAZI, 2011).

3.1.2 *Articulatory*

Os métodos de síntese *Articulatory* tentam fazer uma modelagem dos órgãos vocais humanos, de forma que fique o mais fiel possível. Tendo isso em vista, ele é um dos que mais tem a possibilidade de gerar síntese de voz de alta qualidade (KRöGER, 1992). Essa modelagem é feita criando um modelo sintético da fisiologia do aparelho fonador do homem (PALO, 2006). Alguns dos parâmetros que podem ser considerados neste tipo de sintetização, dependendo do modelo, são: abertura dos lábios, altura da ponta da língua, altura da língua, posição da mandíbula, véu palatino, etc (KRöGER, 1992; PALO, 2006).

Trato Vocal, Acústico, Glottis e Fonte de Ruído são alguns desses modelos que utilizam os métodos de síntese *Articulatory* (KRöGER; BIRKHOLZ, 2009). O primeiro tem como

[2] Consoantes geradas pelo bloqueio da passagem de ar

objetivo a geração de informações geométricas do trato vocal (forma e localização dos órgãos do trato vocal. e.g.: lábios, língua, cavidade nasal) e suas variações ao longo do tempo (SIDDHI; VERGHESE; BHAVIK, 2017). Já o modelo Acústico, segundo Birkholz et al. (2015), tem como função calcular o tempo de variação do fluxo e da distribuição da pressão do ar dentro do trato vocal para calcular o sinal da fala. O Modelo de Glotes visa gerar o sinal da fonte acústica para a fonação e sua inserção no modelo do tubo do trato vocal. Enquanto o Fonte de Ruído gera e insere sinais de fonte de ruído no modelo de linha de transmissão acústica. Os sinais de ruídos resultam de um fluxo de ar turbulento no trato vocal (BIRKHOLZ et al., 2015).

Um dos maiores gargalos da síntese *Articulatory* é a obtenção de dados para o modelo a ser feito. Segundo Klatt (1987), esses dados consistem em uma base, relativamente grande, de movimentos do trato vocal, que pode ser obtida através da análise de raios X da fala humana. Outro ponto é que, de acordo com Kröger (1992), este é um dos métodos mais difíceis de serem implementados, em relação a outros. Uma das dificuldades, por exemplo, é encontrar um balanceamento entre uma boa acurácia e um método fácil de ser mantido e controlado (TABET; BOUGHAZI, 2011). Por essas razões, ele é dos menos utilizados.

3.1.3 *Concatenative*

Uma das principais limitações das sínteses de *Formant* e *Articulatory* é a escolha ou geração de parâmetros de entrada. Essa barreira não é encontrada na síntese *Concatenative*, uma vez que ela utiliza uma abordagem baseada em dados (TABET; BOUGHAZI, 2011). Como pré-requisito, tem-se uma base de dados com unidades de falas, que podem ser fonemas, sílabas, palavras, sentenças, etc. Assim, para formar uma nova fala, o sistema de sintetização escolhe unidades consideradas apropriadas (para aquela nova fala) e faz um processamento de sinal para uní-las. Um dos objetivos desse processamento é que não fique tão perceptível a concatenação de tais unidades (KHAN; CHITODE, 2016).

Um dos métodos de síntese de fala do tipo *Concatenative* é a Seleção de Unidade, onde múltiplas instâncias de unidades de fala com diferentes características prosódicas são armazenadas. No momento em que alguma síntese for feita, a unidade é selecionada do banco de dados com base em dois custos: um custo alvo e um custo de concatenação. Tal tipo de síntese produz falas com boa qualidade em relação à naturalidade, no entanto requer um grande volume de dados (SIDDHI; VERGHESE; BHAVIK, 2017).

Um dos principais desafios da síntese *Concatenative* é em relação ao tamanho da unidade. Uma unidade maior dá maior naturalidade à fala final, pelo fato de não serem necessárias muitas concatenações, no entanto requer uma grande quantidade de memória e é menos flexível. Por outro lado, unidades menores são mais flexíveis e gastam menos memória. No entanto, dependendo de quão pequena a unidade é, sua coleta torna-se complexa e o som final soa menos natural devido uma quantidade maior de concatenações

serem necessárias (LEMMETTY, 1999; THOMAS et al., 2006).

3.1.4 Paramétrico Estatístico

A síntese baseada na abordagem paramétrica estatística não utiliza segmentos de fala naturais. Ela gera fala a partir de modelos estatísticos previamente aprendidos, exigindo menos armazenamento de segmentos naturais, diferentemente dos métodos de concatenação (ZEN; TOKUDA; BLACK, 2009).

O método de síntese de fala baseado em *Hidden Markov Model* (HMM) é a abordagem paramétrica estatística mais comumente utilizada (ZEN; TOKUDA; BLACK, 2009). É composto por duas fases, a de treinamento e a de síntese. Na primeira, os parâmetros de fala são extraídos das entradas do banco de dados de treinamento de fala e são modelados como HMMs. Na segunda fase, as palavras a serem sintetizadas são identificadas no banco de dados e os parâmetros são extraídos desses HMMs. Por fim, a fala é sintetizada a partir desses parâmetros extraídos (RAITIO et al., 2011).

A principal vantagem da síntese de voz baseada em HMM é a flexibilidade, pois a fala é armazenada na forma de parâmetros e é fácil de modificá-los. Já a desvantagem é que as falas produzidas têm grau de naturalidade baixo. Além disso, requer alto processamento devido a necessidade de suavizar os parâmetros (RAITIO et al., 2011). No entanto, Zen, Senior e Schuster (2013) propuseram um método paramétrico estatístico usando redes neurais profundas, onde o custo computacional é menor do que o HMM.

3.1.5 Síntese Híbrida

A síntese híbrida é uma combinação de mais de uma abordagem. Tiomkin et al. (2011) desenvolveram um trabalho utilizando tal método, fazendo uso da síntese *Concatenative* e da estatística paramétrica HMM. Os autores tentaram mitigar alguns pontos negativos da primeira, usando artifícios da segunda e vice-versa.

A *Concatenative* produz falas com boa naturalidade, no entanto quando são utilizados segmentos pequenos para a geração de alguma fala, as fronteiras de concatenação tornam-se mais perceptíveis. Já a estatística paramétrica HMM não permite gerar áudios tão naturais quanto a primeira, mas ela é capaz de suavizar as fronteiras de concatenação da fala, mitigando o problema da descontinuidade e deixando menos perceptível o ponto de concatenação (TIOMKIN et al., 2011).

3.1.6 WaveNet

A abordagem mais atual e poderosa de sintetização de áudios é baseada em WaveNet. Com ela, é possível gerar áudios mais naturais, ou seja, mais parecidos com a voz humana e nos mais variados tipos de idiomas. WaveNet é uma rede neural profunda para a geração de formas de onda de áudios (OORD et al., 2017).

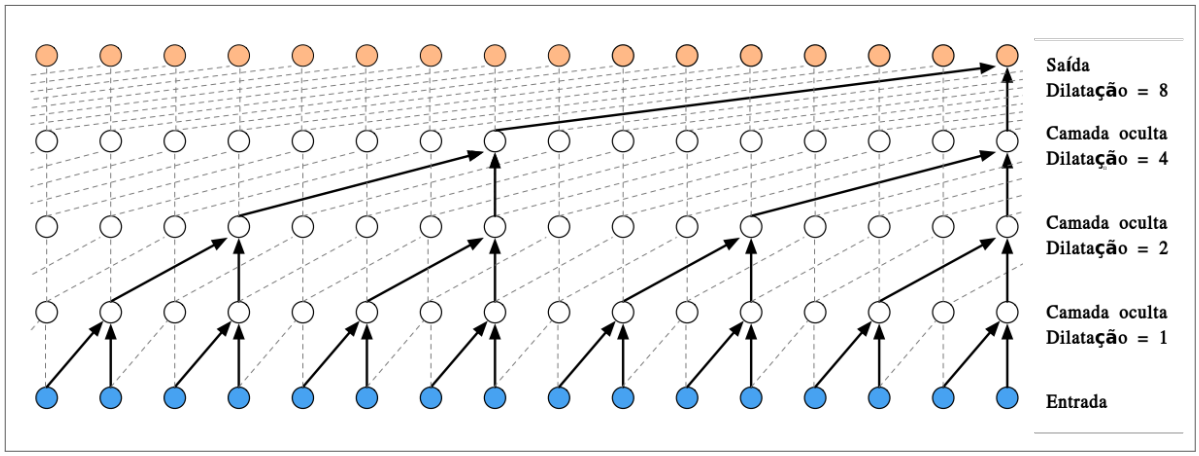


Figura 10 – Visualização de uma pilha WaveNet (OORD et al., 2016)

A Figura 10 mostra como uma WaveNet é estruturada. Ela é uma rede neural convolucional, em que nas camadas convolucionais existem fatores de dilatação que permitem que seu campo receptivo cresça exponencialmente com profundidade e cubra milhares de timesteps. Ela tem como entrada formas de onda reais gravadas por humanos. Após o treinamento, é possível utilizar a rede para gerar áudios sintéticos (OORD et al., 2017).

3.2 SERVIÇOS DE SINTETIZAÇÃO

Algumas empresas fornecem serviços que fazem a sintetização de fala. Uma delas é a Amazon com um serviço chamado Polly[3]. Outra é a Google com uma API chamada *Text to Speech*[4]. Ainda tem a IBM, com o Watson *Text to Speech*[5] e a Microsoft, com o Azure *Text to Speech*[6].

Usando esses serviços, é possível gerar vários áudios de pessoas de diferentes lugares, sotaques e gêneros. Além disso, essa alternativa é consideravelmente barata e os áudios não têm (ou quase não têm) ruído; o que é diferente das abordagens mencionadas nas Seções 2.3.1 e 2.3.2. A seguir, tais abordagens de sínteses serão detalhadas.

3.2.1 Amazon Polly

Polly é um serviço em nuvem para síntese de fala desenvolvido pela Amazon, que tem como objetivo prover falas mais realistas possíveis, utilizando aprendizado profundo para tal. Este serviço oferece a disponibilidade de sintetização em diversos idiomas de diferentes nacionalidades. Para cada idioma, o serviço pode gerar áudios de voz de pessoas diferentes, com gêneros distintos (AWS, 2018). A quantidade de vozes disponíveis atualmente é 53 (33 femininas e 20 masculinas), distribuídas em 27 idiomas.

[3] https://aws.amazon.com/polly/
[4] https://cloud.google.com/text-to-speech/
[5] https://text-to-speech-demo.ng.bluemix.net/
[6] https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/

O Inglês, idioma alvo de avaliação neste trabalho, é o que tem mais vozes disponíveis. Existem opções de vozes para cinco locais diferentes: Austrália, Inglaterra, Índia, Estados Unidos e País de Gales (Tabela 1). Este último só tem opção de voz em um gênero (masculino), enquanto o da Índia, só tem vozes femininas. Já com os demais, é possível sintetizar áudios em qualquer gênero.

No geral, levando em consideração apenas a sintetização de fala em Inglês deste serviço (Polly), é possível sintetizar uma sentença com 16 vozes diferentes, 10 femininas e 6 masculinas. O Americano é o que tem mais opções de vozes, 5 femininas e 3 masculinas, totalizando metade das disponíveis. Na Tabela A.1 que está no Apêndice A, é possível ver com mais detalhes a distribuição das demais vozes dos outros idiomas.

Tabela 1 – Vozes em Inglês Disponíveis no Amazon Polly

Idioma	**Nome(s) Feminino(s)**	**Nome(s) Masculino(s)**
Inglês Australiano	Nicole	Russell
Inglês Britânico	Amy Emma	Brian
Inglês (Índia)	Aditi (bilíngue com hindi) Raveena	
Inglês Americano	Ivy Joanna Kendra Kimberly Salli	Joey Justin Matthew
Inglês Galês		Geraint

Fonte: (AWS, 2018)

O processo de sintetização de fala no Amazon Polly se dá da seguinte forma: dada uma sentença de entrada e nome da pessoa referente às vozes disponíveis, ele produzirá um som de acordo com tais informações. Tal sentença pode ser informada sem formatação (só a string mesmo) ou em *Speech Synthesis Markup Language* (SSML). Desta última forma, também é possível variar alguns parâmetros, personalizando, assim, a saída, a voz sintetizada. Já as sentenças que forem sintetizadas da primeira forma receberão os parâmetros default do sistema.

SSML é uma linguagem de marcação baseada em XML que auxilia na geração da fala sintética. Um dos seus objetivos principais é padronizar uma forma de requisição para a sintetização da fala por determinado serviço, permitindo ao usuário inserir alguns parâmetros como volume, velocidade, pronúncia, etc (W3C, 2018). Nem todos os serviços de sintetização produzirão fala com qualquer um dos parâmetros que essa linguagem

permite, isso vai depender da capacidade e particularidades de cada um. Existem alguns, por exemplo, que só produzem determinados tipos de formatação de saídas para certas vozes.

Exemplos das duas formas de entrada para o sintetizador estão na Tabela 2. A primeira, é o tipo de entrada padrão, uma sentença sem informações adicionais. Ela, por sua vez, será sintetizada de acordo com os parâmetros pré-definidos pelo sistema. Enquanto a segunda, está no formato SSML. No exemplo dado, o volume da sentença a ser sintetizada será mais alto que o volume default do Polly.

Tabela 2 – Exemplos de Entrada para o Polly

1	**O volume está com altura adequada.**
2	<speak><prosody volume='loud'> </prosody>**O volume está alto.**</speak>

Fonte: Este trabalho

Além do volume, outras personalizações que podem ser feitas na síntese, quando se usa SSML neste serviço, são a respeito da velocidade e do timbre da fala. Também é possível incluir pausas no meio do texto de entrada. Se esse texto for consideravelmente grande, pode-se incluir respiração, dando maior naturalidade ao o que foi sintetizado. Há também a possibilidade de enfatizar determinada palavra ou trecho da sentença, e até mesmo de sussurrar.

Quanto ao formato da saída, há a opção do usuário escolher. É possível gerar áudios com extensões mp3, ogg e vorbis. A escolha do formato de áudio vai depender da aplicação.

3.2.2 Google *Text to Speech*

A Google também lançou um serviço de sintetização de fala, que é chamado de *Text to Speech* (GOOGLE, 2018). Da mesma forma que o Polly, é possível sintetizar fala em várias línguas e diferentes tipos de pessoas. As vozes produzidas pelo Google TTS podem ser Padrão ou WaveNet. Uma fala Padrão também pode ser chamanda por Não WaveNet, uma vez que ela não foi sinteitzada utilizando tal abordagem. Pode ter sido utilizando métodos concatenativos e/ou estatístico paramétricos (GONZALVO et al., 2016; ZEN et al., 2016). Falas geradas utilizando a abordagem WaveNet têm melhor qualidade do que as que utilizam essas abordagens tradicionais (OORD et al., 2016; SHEN et al., 2018).

O número de idiomas é menor que o do Amazon Polly, que tem 26, enquanto o da Google tem 14. No entanto, a quantidade de vozes diferentes é superior, que são 56 do Google TTS (34 femininas e 22 masculinas) e 53 do Polly. Ou seja, a cobertura em número de idiomas do serviço da Amazon é maior, já dentro dos idiomas disponíveis, a cobertura

do número de vozes para cada um, no Google é maior. A distribuição dessas vozes nos idiomas estão detalhadas no apêndice A, Tabela A.2.

Igual ao da Amazon, neste serviço, o Inglês é o idioma que tem mais opções de vozes. Na Tabela 3 pode-se observar todas as combinações possíveis. Embora o número de locais diferentes sejam apenas 3 (Austrália, Inglaterra e Estados Unidos), menos que os da Polly, a quantidade de vozes em cada idioma é maior. O Americano tem 10, sendo 5 masculinas e 5 femininas e os outros dois, 8 vozes, 4 de cada gênero. Assim, são 26 vozes diferentes para gerar áudios em Inglês. Dessas 26, 16 são consideradas padrão e as demais, Wavenet, as ditas com maior qualidade.

Tabela 3 – Vozes em Inglês Disponíveis no Google *Text to Speech*

Idioma	**Tipo de Voz**	**Vozes Femininas**	**Vozes Masculinas**
Inglês Australiano	Padrão	en-AU-Standard-A	en-AU-Standard-B
		en-AU-Standard-C	en-AU-Standard-D
	WaveNet	en-AU-Wavenet-A	en-AU-Wavenet-B
		en-AU-Wavenet-C	en-AU-Wavenet-D
Inglês Britânico	Padrão	en-GB-Standard-A	en-GB-Standard-B
		en-GB-Standard-C	en-GB-Standard-D
	WaveNet	en-GB-Wavenet-A	en-GB-Wavenet-B
		en-GB-Wavenet-C	en-GB-Wavenet-D
Inglês Americano	Padrão	en-US-Standard-C	en-US-Standard-B
		en-US-Standard-E	en-US-Standard-D
	WaveNet	en-US-Wavenet-C	en-US-Wavenet-A
		en-US-Wavenet-E	en-US-Wavenet-B
		en-US-Wavenet-F	en-US-Wavenet-D

Fonte: (GOOGLE, 2018)

A respeito do processo de sintetização, dado um texto de entrada e o nome da voz, é possível produzir um áudio, de acordo com tais parâmetros. Esse texto pode ser no formato padrão ou dentro de tags SSML, o que dá maior flexibilidade de mudar algumas características do áudio a ser produzido. Tais características são: volume, velocidade, timbre, dentre outras.

A Tabela 4 mostra dois tipos de exemplos de entrada, um que seria o padrão e outro, no formato SSML. O texto de entrada a ser sintetizado é o número 5. No primeiro caso, será gerado um áudio com uma voz falando "cinco". No segundo caso, foi utilizada a tag "say-as", com o parâmetro "interpret-as=’ordinal’", ou seja, o que se pede é que seja gerada um áudio com o número 5 dito de forma ordinal. Assim, será gerado um som falando "quinto".

Tabela 4 – Exemplos de Entrada para o Google TTS

1	**5**
2	<speak><say-as interpret-as='ordinal'> **5**</speak>

Fonte: Este trabalho

Os áudios podem ser produzidos nos formatos mp3, ogg e Linear16.

3.2.3 IBM Watson *Text to Speech*

O serviço de sintetização da fala da IBM, conforme visto em Pitrelli et al. (2006), utiliza métodos concatenativos para fazer a síntese. O uso de tal abordagem é confirmado em trabalhos que foram publicados algum tempo depois por pesquisadores da IBM. Em um deles, desenvolvido por Fernandez et al. (2016), foi explorado o uso de Redes Neurais Recorrentes Bidirecionais como um modelo de predição de prosódia dentro de um sistema TTS de seleção de unidade (um dos métodos de síntese *Concatenative*). Em outro, feito por Sorin, Shechtman e Rendel (2017), foi desenvolvido um sistema TTS concatenativo com recursos de transformação instantânea de voz. Esses trabalhos fazem parte da composição do serviço de síntese da fala da empresa em questão que é chamado de Watson TTS (IBM, 2018; RESEARCH-HAIFA, 2019).

Utilizando o IBM Watson TTS também é possível gerar falas de mais de uma pessoa, no entanto, endo como referência os serviços da Amazon e da Google, a quantidade de vozes disponíveis no Watson é pequena. São apenas 14 (11 femininas e 3 masculinas), enquanto cada um dos outros têm mais de 50. O número de idiomas também é reduzido, são apenas 10. Dessa forma, tanto a cobertura do número de idiomas, quanto das vozes para cada idioma, é consideravelmente baixa, principalmente quando se fala em voz masculina. Apenas em três idiomas é possível produzir voz sintetizada masculina, Alemão, Espanhol Casteliano e Inglês Americano. É possível ver essas informações em detalhes na tabela 15 no apêndice A.3.

Juntamente com o Espanhol, o Inglês é o idioma que mais tem vozes disponíveis pra sintetização utilizando este serviço. Há dois tipos de Inglês, o Americano e o Britânico, com 1 e 3 vozes, respectivamente. Destas quatro, apenas uma é masculina, a de Michael com idioma dos Estados Unidos. É possível observar tais informações detalhadamente na tabela 5.

Tabela 5 – Vozes em Inglês Disponíveis na IBM Watson

Idioma	**Voz(es) Feminina(s)**	**Voz(es) Masculina(s)**
Inglês Britânico	Kate	
Inglês Americano	Alisson	Michael
	Lisa	

Fonte: (IBM, 2018)

O formato de entrada em SSML também é aceito nesse tipo de aplicação. Sendo assim, possível, customizar o áudio de saída. Na Tabela 6 tem dois exemplos de entrada, um padrão e outro, usado a linguagem de marcação. No primeiro exemplo, o áudio de saída será o default, sem nenhuma expressão. Já no segundo, a tag "express-as"com o parâmetro "type=’GoodNews’"indica a fala apresentará certa emoção, um entusiasmo, quando se dá uma boa notícia. Outras emoções que podem ser passadas como parâmetros são de desculpa e incerteza. Esse tipo de customização, no entanto, só é disponível para uma das vozes.

Tabela 6 – Exemplos de Entrada para o IBM Watson

1	**Consegui uma boa nota na prova.**
2	<speak><express-as type=’GoodNews’> **Consegui uma boa nota na prova.**</speak>

Fonte: Este trabalho

Alguns formatos de áudios que podem ser gerados pelo serviço da IBM são: Ogg, Vorbis, WAV, FLAC, MP3, l16 (PCM) e mulaw. Tendo, assim, mais opções que os serviços da Amazon e da Google.

3.2.4 Microsoft Azure *Text to Speech*

Outra empresa que tem investido em sintetização da fala é a Microsoft, com uma API da Azure (MICROSOFT, 2018). As falas são sintetizadas utilizando a abordagem estatística paramétrica e/ou concatenativa (ACERO; HON; HUANG, 1998; LING; DENG; YU, 2013; CHANDU et al., 2017). Dos quatro serviços de conversão do texto em fala abordados neste trabalho, este é o que dá suporte a mais idiomas e localidades, totalizando 46. A quantidade de vozes também é maior do que nos outros serviços, 46 vozes femininas e 35 masculinas.

Tratando-se do Inglês, é o idioma que tem mais opções de vozes também neste serviço. O total é de 16, dividido por 6 localidades: Canadá, Inglaterra, Irlanda, Índia, Estados

Unidos e Austrália. Dessas 16, apenas 5 são masculinas, consequentemente não cobre todas as localidades. Assim, não é possível sintetizar fala masculina com Inglês Canadense, nem Australiano. Já com as femininas, é possível sintetizar fala para todos os tipos de Inglês, até com mais de uma opção, exceto o da Irlanda, que só gera áudio do gênero masculino. As informações estão detalhadas na tabela 7. Esse serviço também dá suporte à entrada tanto de texto sem formatação, quanto SSML. Algumas das customizações possíveis de serem feitas utilizando este serviço são: de volume, densidade, velocidade, pronúncia e quebra da frase.

Tabela 7 – Vozes em Inglês Disponíveis no Microsoft Azure

Idioma	**Nome(s) Feminino(s)**	**Nome(s) Masculino(s)**
Inglês Canadense	Linda HeatherRUS	
Inglês Britânico	Susan HazelRUS	George
Inglês (Irlanda)		Sean
Inglês (Índia)	Heera PriyaRUS	Ravi
Inglês Americano	ZiraRUS JessaRUS Jessa24kRUS	BenjaminRUS Guy24kRUS
Inglês Australiano	Catherine HayleyRUS	

Fonte: (MICROSOFT, 2018)

Na Tabela 8 pode-se ver exemplos do do tipo de entrada sem formatação, o primeiro, e tipo de entrada em formato SSML, o segundo. No primeiro caso, a fala a ser gerada seguirá as configurações padrão do sistema, sem alteração nenhuma alteração. Já no segundo, ocorrerá uma quebra, pausa no meio da fala. Isso é possível por causa da tag "break"com o parâmetro "time=’100ms’".

Tabela 8 – Exemplos de Entrada para o Microsoft Azure TTS

1	**Este texto não tem quebra**
2	<speak>**Texto com**<break time=’100ms’ />**quebra.**</speak>

Fonte: Este trabalho

Diferentemente de outros serviços que podem gerar áudios em mais de um tipo de formato, a Azure TTS produz fala apenas no formato Wav. Caso alguém opte por usar este serviço e deseja outra extensão, tem que fazer a conversão para o formato desejado.

3.3 CONSIDERAÇÕES FINAIS

Nesse capítulo, foram apresentados os quatro métodos tradicionais de sintetização encontrados na literatura, *Formant, Articulatory, Concatenative* e *WaveNet*. A utilização deste último vem crescendo atualmente, devido a sua capacidade de gerar áudios mais naturais. A Google deixa claro que em seu serviço de sintetização, algumas das falas geradas usando tal método.

Além dos serviço da Google, foram apresentados outros três mais comuns no mercado. Uma vez que haja a necessidade de fazer uso de algum, a primeira etapa é verificar se o serviço suporta o idioma desejado, eliminando assim, os que não tem. Conforme visto, a cobertura de idiomas é maior no Microsoft Azure, seguindo do Amazon Polly, Google TTS e IBM Watson.

Outro fator que é levado em consideração é a naturalidade da voz produzida pelo sistema, o quanto mais parecida com a humana ela é. Além disso, tem a questão de variedade de vozes dentro de um mesmo idioma. O serviço da Google é o que tem mais variações em determinados idiomas. Isso pode ser visto no apêndice A.

A possibilidade que o usuário tem de customizar a voz de saída também é um fator a ser levado em consideração. Como todos os serviços aceitam o formato de entrada em SSML, então aceitam customização, no entanto existem alguns mais restritos. Os mais adaptáveis são os da Amazon e da Google.

4 ARQUITETURA DE TESTE DE SISTEMAS *SPEECH TO TEXT*

Neste capítulo, será apresentada a arquitetura para teste automático de sistemas STT proposta por este trabalho. Na Seção 4.1, será detalhado como se dá o processo desse tipo de teste em um ambiente real, nos smartphones da Motorola Mobility. Assim, o problema será minuciado e em seguida, na Seção 4.2, a arquitetura proposta que visa mitigá-lo será descrita. Por fim, na Seção 4.3, serão descritos os detalhes de implementação.

4.1 CONTEXTO GERAL E OBJETIVO

O desenvolvimento deste trabalho foi feito em um ambiente real de teste de sistema STT em *smartphones*, em parceria com a Motorola Mobility. Para simplificar, esse tipo de teste será chamado de "teste de voz". Tais testes são feitos em um aplicativo chamando de Moto Voice que incorpora modelos diferentes de transcrição da fala e interpretação dos comandos falados, para o possível execução da tarefa referente.

O Moto Voice escuta comandos de voz ditos pelo usuário e envia para o modelo A de transcrição. Caso consiga transcrever e interpretar o comando, ele executa a tarefa referente ao que foi entendido. Caso A não consiga transcrever, o comando é enviado para o modelo B de transcrição que também vai interpretá-lo e retornar o resultado para o usuário. Se B fizer a transcrição, mas não a interpreta, ela é enviada para um modelo X que vai tentar interpretá-la e enviar a resposta. No entanto, se B não conseguir ao menos transcrever, o comando de voz é enviado para o modelo C de transcrição e depois enviado para X interpretá-la e executar a tarefa correspondente.

Antes das etapas descritas no parágrafo anterior, é necessário que o Moto Voice seja ativado. Tal ativação também pode ser via voz. Para isso, é necessário que o usuário grave um comando que sirva de chave de ativação sempre que quiser usar o aplicativo. Só depois de falar o "comando-chave", ele pode falar/perguntar o que deseja que seja executado (no celular).

Assim como outros tipos de teste de software, o de voz pode ser feito de forma manual ou automática. Na primeira forma, é necessário um testador (humano) falar diretamente comandos e verificar a resposta dada pelo aparelho. Neste processo, é verificado se o que o testador falou foi entendido corretamente (*i.e.*, se o sistema fez a transcrição correta da fala para texto). Além disso, o testador também observa se depois de transcrito, o sistema dará a resposta certa à solicitação do usuário. Esse tipo de teste manual é repetitivo, fatigante e custoso.

Visando diminuir o trabalho manual, tem-se os testes que podem ser executados automaticamente. Ao invés de terem pessoas executando, pode-se fazer implementação de scripts de teste que serão executados por uma máquina quantas vezes forem necessárias.

Da mesma forma que nos executados manualmente, nos automáticos também são verificadas as transcrições e a resposta dada pelo sistema. No entanto, diferente de outros tipos de testes automáticos de software, os de voz requerem um tipo de artefato adicional, que são os comandos de voz em forma de áudio. A aquisição dos áudios usados como casos de teste tem sido um dos principais gargalos desse processo de teste.

A aquisição dos áudios para fazer os testes automáticos de voz nos aparelhos da empresa em questão tem sido feita de forma manual, i.e., através de gravação de comandos por humanos. A princípio, tem-se uma lista de comandos escritos por um profissional da área de teste, que precisam ser gravados. Para a gravação é necessário que a pessoa esteja em um ambiente com pouco ruído e usando equipamentos adequados, como um bom microfone, para que os áudios tenham melhor qualidade. Depois disso, tais áudios passam por um processo de verificação manual (pós-processamento), onde outro profissional vai escutá-los, e aqueles que são considerados com qualidade ruim são descartados. Esse processo é custoso, tanto em questão de tempo, quanto financeira.

Segundo um engenheiro de teste da Motorola que trabalha com testes automáticos de voz, uma pessoa leva em torno de 10 minutos para a gravação de um áudio, sem contar com a etapa de pós-processamento. Esse tempo gasto é relativamente grande por causa da necessidade do *setup* do ambiente, equipamentos e ferramentas para a gravação. Levando em consideração que para alguns tipos de teste, como os de acurácia, centenas de comandos são necessários e gravados por pessoas diferentes, de diversas nacionalidades e sotaques, esse custo de tempo torna-se muito alto.

Para adquirir áudios com tal variedade, é necessário ir a países diferentes e contratar pessoas para fazer as gravações. Ou seja, dado um comando A em inglês, ele precisa ser gravado por pessoas de países onde o inglês é língua oficial, como Estados Unidos, Austrália, Canadá, Inglaterra, e outros. O mesmo se aplica a comandos em outras línguas. Isso gera um custo financeiro também alto.

Além desta questão de aquisição das falas com diferentes variações para dar maior cobertura aos testes, ainda existe outro fator responsável para garantir a robustez do software, que são testes com variações de um dado comando de entrada. Tal variação é em relação à forma de falar e à estrutura sintática. Um usuário pode, por exemplo, solicitar algo ao (interagir com) sistema de diferentes maneiras, esperando a mesma resposta. Ele pode perguntar a previsão do tempo de diversas formas, e é importante que o sistema dê a resposta certa independentemente de como foi perguntado, desde que faça sentido. Nesse contexto, surge outro gargalo que é a necessidade de gerar possíveis variações dos comandos previamente escritos.

Assim, este trabalho propõe uma solução a fim de mitigar estes impasses, tanto no processo de aquisição de áudios para execução dos testes automáticos de voz, quanto na geração automática de sentenças equivalentes. Sobre a aquisição dos áudios, tendo em vista que o principal fator do problema atual é a necessidade de ter várias pessoas, de

diferentes gêneros e sotaques gravando os comandos, foi proposta uma arquitetura em que não são necessárias pessoas para fazer a gravação das falas. Ao invés disso, são produzidas falas sintetizadas.

Para a geração de tais falas de forma automática foram utilizados os quatro maiores serviços de sintetização do mercado, que foram introduzidos no Capítulo 3, Seção 3.2. Depois, foram feitos experimentos a fim de verificar se os áudios humanos poderiam ser substituídos pelos sintetizados nos testes de voz e até que ponto isso seria possível. Isso será mostrado no Capítulo 5.

4.2 ARQUITETURA PROPOSTA

A arquitetura proposta é dividida em cinco módulos principais: pré-processamento de sentenças (1), sintetização da fala (2), geração de sentenças equivalentes (3), filtragem de sentenças (4) e teste do sistema STT (5). Na Figura 11, está ilustrado como se dá a comunicação entre tais módulos e na Figura 12, o fluxo geral.

A entrada para o sistema é uma lista de sentenças em inglês, que será recebida pelo módulo de pré-processamento, que faz a identificação de sentenças interrogativas e a remoção daquelas que são repetidas. O resultado é uma nova lista que servirá de entrada para o módulo de sentenças equivalentes. Em tal módulo (2), uma lista de sentenças equivalentes a cada uma das presentes da lista de entrada, será gerada. Dada uma sentença A, presente na lista, um conjunto X de sentenças sintaticamente similares será gerado, de acordo com os sinônimos de cada palavra presente no texto de entrada. Para isso, é utilizado quatro alternativas diferentes: O WordNet[1]; os dicionários Collins[2], Oxford[3] e Thesaurus[4]. O que resulta da etapa 2 servirá de entrada para o módulo 3, de filtragem.

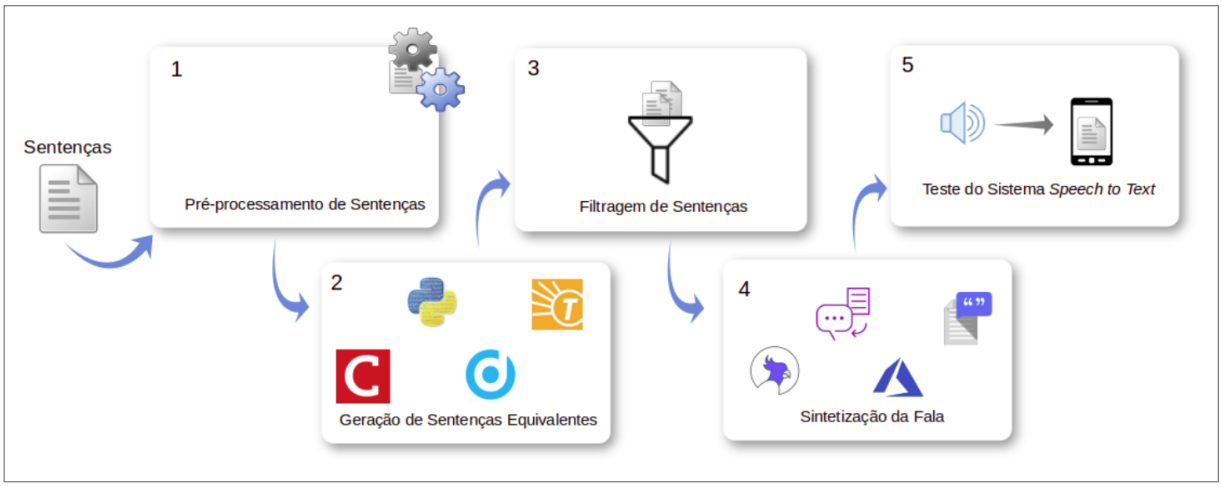


Figura 11 – Arquitetura Proposta

[1] http://www.nltk.org/howto/wordnet.html
[2] https://www.collinsdictionary.com
[3] https://en.oxforddictionaries.com
[4] https://www.thesaurus.com

O conjunto de sentenças resultantes da etapa 2 pode ser consideravelmente grande, devido a alta quantidade de sinônimos, dependendo da palavra. Para fazer a redução da quantidade de sentenças geradas, existe o módulo de filtragem, que aplica uma métrica de similaridade de sentenças, eliminando as que estão abaixo de um limiar de similaridade; e ainda implementa uma abordagem baseada em sinônimos (detalhes serão vistos na Seção 4.3.3). O resultado desta etapa é uma lista de sentenças filtrada que serve como entrada para o módulo de sintetização da fala.

No módulo 4 tais sentenças serão transformadas em áudios, de acordo com o serviço que o usuário escolher, se é o da Amazon, Google, IBM ou Microsoft (detalhados na Seção 3.2). A lista de sentenças pode ser em formato normal, só o texto mesmo, ou em SSML. Esta segunda forma dá ao usuário maior liberdade de colocar mais parâmetros, desde que sejam permitidos pelo serviço escolhido. Além da lista de sentenças em SSML, ou não, e do serviço, o usuário também tem a liberdade de informar a voz que será usada, o gênero e o inglês referente a algum local (EUA, Inglaterra, Canadá), dependendo do que for oferecido pelos serviços. Depois disso, os áudios gerados servirão como entrada para a execução dos testes no sistema.

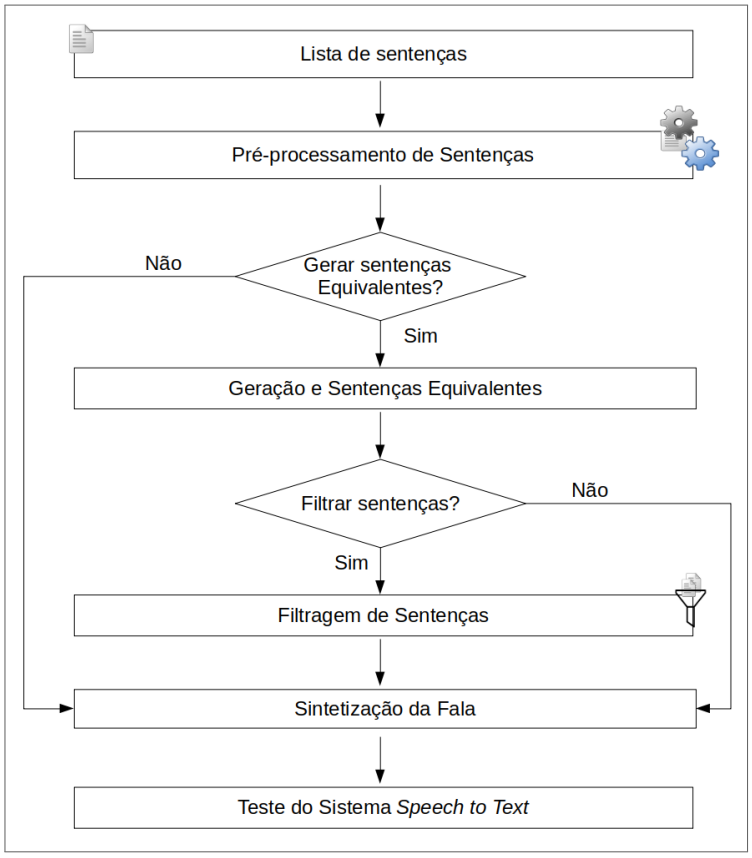


Figura 12 – Fluxo Geral da Arquitetura Proposta

O último módulo (5), é o de teste propriamente dito do sistema de reconhecimento da fala. Nele, é verificado, de forma automática, se esse software, que está em um celular, consegue reconhecer corretamente aquilo que foi reproduzido pelas falas sintetizados.

Também é verificada a resposta dada pelo sistema à requisição referente ao comando falado.

Esses são os módulos fundamentais da arquitetura para teste de sistemas STT proposta por este trabalho. Na seção 4.3, os detalhes de implementação deles serão descritos.

4.3 IMPLEMENTAÇÃO

A arquitetura proposta é dividida em cinco módulos, os vistos na Seção 4.2, que serão detalhados nesta. Tais módulos são: Pré-processamento de Sentenças, Sintetização da Fala, Geração de Sentenças Equivalentes, Filtragem de Sentenças e Teste de Sistema *Speech to Text*.

4.3.1 Pré-processamento das Sentenças

Embora seja o primeiro módulo da arquitetura proposta, não foi o primeiro a ser implementado. Seu desenvolvimento surgiu da identificação de problemas durante o uso de tal arquitetura. Foi observado que, em alguns casos, o usuário disponibilizava listas de sentenças relativamente grandes, com centenas de comandos, com erros que precisavam ser corrigidos. Alguns destes erros eram: comandos repetidos, sem padrão de escrita e sentenças interrogativas sem o sinal de interrogação. Assim, foi observada a necessidade de fazer um pré-processamento das sentenças de entrada antes de serem usadas para a fase de sintetização em si. As etapas de processamento de sentenças são mostradas na Figura 13.

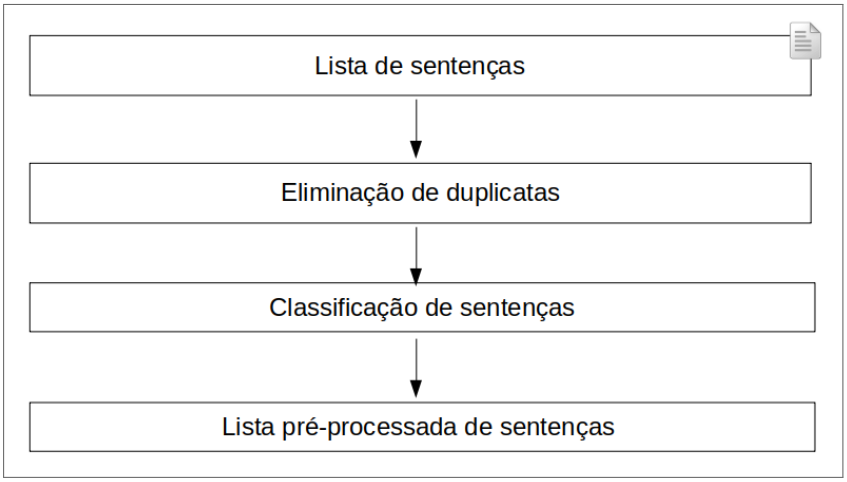


Figura 13 – Pré-processamento da Lista de Sentenças

Ter comando repetido na lista não é interessante porque ele será sintetizado mais de uma vez, sem que haja necessidade, aumentando assim a complexidade de tempo da sintetização como todo. Desta forma, há uma verificação se há sentenças repetidas na lista, se sim, elas são eliminadas, ficando somente uma de cada.

Outro fator é que frases interrogativas, no geral, tem uma entonação diferente das afirmativas e exclamativas. Ter, na lista, uma sentença que é interrogativa sem a pontuação impacta na qualidade da sintetização, uma vez que a fala sintetizada não terá a entonação adequada. Ela pronunciará uma interrogação como se fosse afirmação e isso dá menor naturalidade à fala. Assim, foi usado um método de classificação de sentenças, onde aquelas que estão sem pontuação passam por um classificador responsável por julgar se ela é interrogativa ou não. Por outro lado, se a sentença já pontuada na lista de entrada, esta é a pontuação considerada, não se usa o classificador nesses casos.

Foram testados dois classificadores (treinados com um corpus[5] do *Natural Language Toolkit* (NLTK) que contém frases e suas respectivas pontuações), *Naive bayes* e *Support vector machine* (SVM), respectivamente. Um classificador Naive Bayes é baseado na aplicação do teorema de Bayes com forte independência entre as características. Ele assume que a presença de uma característica particular de uma classe não está relacionada à presença de qualquer outra característica. Já o SVM, é um classificador discriminativo formalmente definido por um hiperplano que separa os dados entre duas classes (VIJAYARANI; MR.S.DHAYANAND, 2015).

Afim de escolher um desses dois classificadores para compor a arquitetura proposta, foram realizados experimentos, onde os dois foram testados utilizando validação cruzada de 30-*fold*. O primeiro a ser testado NLTK teve acurácia em torno de 67%. Já o segundo, se mostrou com melhor desempenho, 98% de taxa de acerto. Desta forma, esse último foi o escolhido para fazer a classificação de sentenças neste trabalho.

4.3.2 Geração de Sentenças Equivalentes

Ao testar um sistema de reconhecimento de fala, deve-se levar em consideração as diferentes formas que os usuários podem interagir. Um software com boa qualidade é capaz de reagir corretamente a essa diversidade de interações. Uma pessoa pode, por exemplo, solicitar algo ao sistema de formas diferentes, mas esperando o mesmo retorno (e.g., ela pode perguntar a previsão do tempo de diversas maneiras). Assim, este módulo surge neste contexto, da necessidade de aumentar a cobertura dos testes, com a finalidade de prover um sistema de reconhecimento de fala mais robusto.

Uma forma de testar o sistema com diferentes entradas, mas esperando o mesmo resultado, é fazendo variações de uma mesma sentença, de forma que o sentido semântico não seja perdido. Fazer isso manualmente é relativamente custoso, assim foi proposta uma maneira de fazê-lo de forma automática, baseada em sinônimos das palavras que compõe a sentença em questão.

Para tal, foram utilizados quatro meios diferentes de aquisição de sinônimos: Wordnet, Oxford, Collins e Thesaurus. De todos os 4 meios citados, no Wordnet é onde são encontrados mais sinônimos, o que possibilita a geração de mais sentenças equivalentes.

[5] https://www.nltk.org/api/nltk.corpus.reader.html#nltk.corpus.reader.xmldocs.XMLCorpusReader

Em contrapartida, dependendo da palavra, boa parte dos sinônimos retornados não fazem muito sentido, tem muito ruído (um exemplo disso é um dos sinônimos retornados para a palavra "lua", "mês lunar"). Oxford, Collins e Theasurus são dicionários, onde para a coleta dos sinônimos dele foram implementados crawlers, para fazer busca na web.

O processo é feito da seguinte forma, conforme ilustrado na Figura 14: dada uma sentença, o sistema faz a quebra dela em palavras e faz buscas de todas as palavras em todas as fontes descritas no parágrafo anterior. Em seguida, para cada fonte, é feita a combinação de todos os sinônimos das palavras que compõe a sentença e depois, todas as frases são armazenadas em apenas um lista.

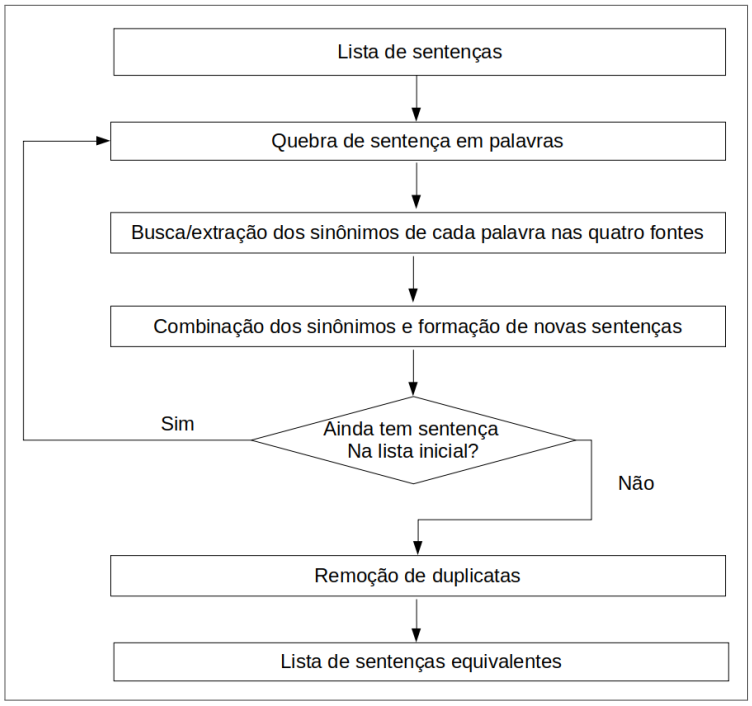


Figura 14 – Geração de Sentenças Equivalentes

A Tabela 9 mostra o exemplo de geração de sentenças equivalentes da frase "não perturbe"pelas buscas feitas em uma das quatro fontes. A sentença dada foi quebrada em duas palavras ("não"e "perturbe") e os sinônimos de cada uma foram retornados e mostrados nas colunas 2 e 3. Em seguida, eles foram combinados (todos os sinônimos de "não", com todos de "perturbe") para fazer a composição da lista final. Tal processo é feito pra cada uma das quatro fontes e todas as sentenças são colocadas em uma mesma lista no final.

Como na lista resultante da etapa anterior tem sentenças oriundas de quatro fontes diferentes, consequentemente podem ter algumas repetidas. Assim é feita uma remoção de sentenças duplicadas antes do resultado final ser retornado para o usuário.

Tabela 9 – Geração de Sentenças Equivalentes

Sentença Incial	Sinônimos de "Não"	Sinônimos de "Perturbe"	Sentenças Equivalentes
Não Perturbe!	Nunca, Negação, Resusa, Jamais	Incomode, Importune, Atrapalhe, Interrompa	Nunca Incomode! Nunca Importune! Nunca Atrapalhe! Nunca Interrompa! . . . Jamais Interrompa!

Fonte: Este trabalho

4.3.3 Filtragem de Sentenças

A lista de sentenças equivalentes resultante do módulo anterior pode ser consideravelmente grande e é possível que sem sempre seja do interesse do usuário usá-la por completo para geração dos áudios. É possível ainda que o usuário use apenas as sentenças equivalentes consideradas mais relevantes. Tendo isso em vista, este módulo, nomeado de filtragem de sentenças, foi desenvolvido (Figura 15).

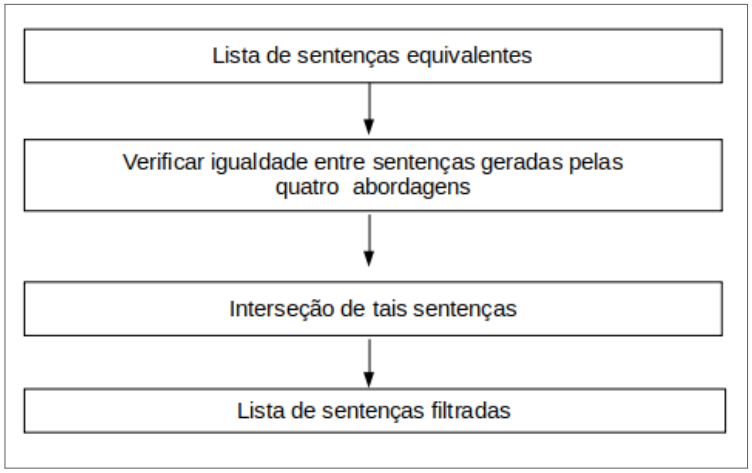


Figura 15 – Filtragem de Sentenças Equivalentes

Para fazer a filtragem, foi feita a interseção das sentenças resultantes daquelas quatro maneiras de geração. Uma determinada sentença que foi gerada das quatro formas, será considerada na etapa de filtragem, ou seja, vai para a lista final. A Tabela 10 mostra o exemplo de tal filtragem.

Tendo a sentença "Não Perturbe!"e suas equivalentes que foram geradas pelas quatro fontes na etapa de geração de sentenças equivalentes, na lista final resultante da etapa de filtragem, foram consideradas apenas três ("Não Incomode!", "Não Atrapalhe"e "Não

Tabela 10 – Filtragem de Sentenças

Sentença Incial	Sinônimos Wordnet	Sinônimos Oxford	Sinônimos Collins	Sinônimos Thesaurus	Lista final de Sentenças
Não Perturbe!	Nunca Incomode! Nunca Importune! Nunca Atrapalhe! Nunca Interrompa! Jamais Interrompa Negação Incomode! . . . Negação Atrapalhe!	Nunca Incomode! Nunca Importune! Nunca Atrapalhe! Nunca Interrompa! Jamais Interrompa! . . . Jamais Interrompa!	Nunca Incomode! Nunca Importune! Nunca Atrapalhe! . . . Nunca Interrompa!	Nunca Incomode! Nunca Atrapalhe! Nunca Interrompa! . . . Jamais Interrompa!	Nunca Incomode! Nunca Atrapalhe! Nunca Interrompa!

Fonte: Este trabalho

Interrompa!"). Essas três são justamente as que se repetem nos quatro grupos de sentenças equivalentes (colunas de 2 a 4).

4.3.4 Sintetização da Fala

Neste módulo, foram feitas implementações com quatro serviços de sintetização diferentes: Amazon Polly, Google TTS, IBM Watson TTS e Microsoft Azure TTS. Foi dada uma visão geral de tais serviços na Seção 3.2. O pré-requisito para utilizá-los é ter credencias para acesso, que podem ser feitas nos sites da respectivas aplicações. Tendo isso, as implementações podem ser feitas.

A função principal de sintetização da fala pode receber como entrada uma lista de sentenças, o serviço, a linguagem, o gênero e o locutor. Destes, apenas o serviço e as sentenças têm que ser obrigatoriamente informados. Ao ser executada, dependendo do serviço escolhido, uma função diferente será chamada, uma vez que cada serviço tem suas particularidades (e.g: tipo de linguagem, nome do locutor). Para melhor entendimento do que ocorre se algum dos parâmetros adicionais (linguagem, gênero e locutor) forem escolhidos, observe a Figura 16.

Se apenas a lista de sentenças e o locutor forem informados, serão gerados áudios referentes somente a ele, de acordo com as sentenças passadas. Assim, se forem passadas 10 sentenças, o serviço for Amazon Polly, com o locutor "Michael", por exemplo, o resultado será 10 áudios com Michael falando (ver Tabela 1, com as vozes em inglês do polly).

Por outro lado, se apenas a lista de sentenças for informada, ou seja, se os demais parâmetros forem vazios, serão gerados áudios disponíveis no serviço referentes a todas os locais onde o inglês é língua oficial. Assim, se o serviço for Amazon Polly, os áudios gerados para cada frase serão todos referentes aos nomes listados na Tabela 1. Se forem passadas as mesmas 10 sentenças do exemplo anterior, 160 áudios serão gerados, 10 para cada um dos 16 locutores.

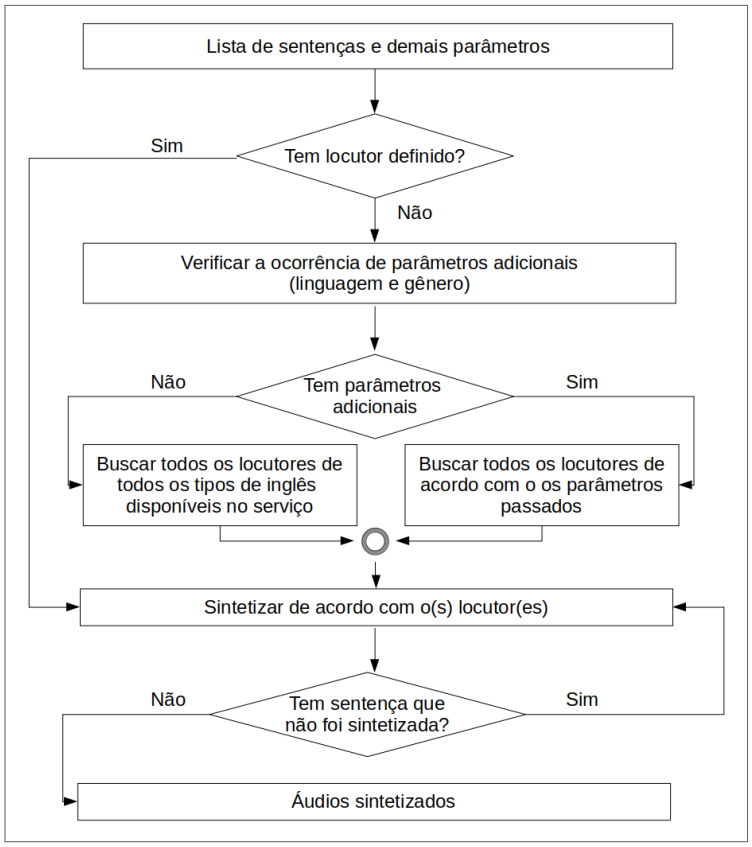


Figura 16 – Sintetização da Fala por Determinado Serviço

Se ao invés do locutor, a linguagem for dita, o que resulta são falas sintetizadas referentes a todos os locutores daquela linguagem. Assim, se o Inglês escolhido for o Britânico do Amazon Polly, as vozes de Ammy, Emma e Brian serão usadas (Tabela 1). Quando, além da linguagem, o usuário informa o gênero, apenas áudios referentes a aquele gênero daquela linguagem serão gerados. Supondo que seja escolhido o gênero feminino também do inglês Britânico do Polly, apenas as vozes de Ammy e Emma serão selecionadas. E, por fim, se apenas o gênero for o parâmetro passado, todos os locutores daquele gênero do inglês de todas as localidades serão usados.

Os áudios que resultam da função de sintetização são em extensão wav (do inglês *WAVEform audio format*) e são organizados da seguinte forma: Tem uma pasta raiz, dentro dela tem outras referentes a cada serviço (no caso, são 4), dentro de cada uma delas tem n pastas referentes às sentenças passadas e dentro das pastas de cada sentença tem os áudios. Cada um deles são nomeados conforme os parâmetros do mesmo (locutor, linguagem, sentença), ficando assim mais fácil de acessá-los e identificá-los.

4.3.5 Teste de Sistema *Speech to Text*

Esta é a última etapa da arquitetura proposta, onde os testes são executados automaticamente com as falas sintetizadas. O sistema a ser testado está em um celular, que, por sua vez, é conectado a um computador com os scripts de teste. No computador, também

é conectada uma caixa de som que fica a certa distância do celular.

Os scripts têm dois papeis fundamentais: por um lado, começar a execução de determinado teste e fazer o áudio referente reproduzir; e por outro, verificar a transcrição daquele áudio pelo sistema STT do celular. Ele faz isso porque já tem conhecimento prévio do que está sendo falado naquele áudio, assim, ele pode fazer a comparação com a transcrição feita pelo software de reconhecimento de fala. Se o sistema transcrever corretamente, então o teste é bem sucedido, senão, ele falha.

A Figura 17 mostra o fluxo de execução deste módulo. Dada uma lista de áudios com a informação dos comandos que estão em cada um deles, o script de teste vai percorrendo tal lista, pega cada áudio e recupera o comando que foi gravado por ele. Depois, reproduz o áudio e verifica no aparelho se a transcrição feita é igual ao comando. No final da execução, um arquivo de log com os resultados é gerado.

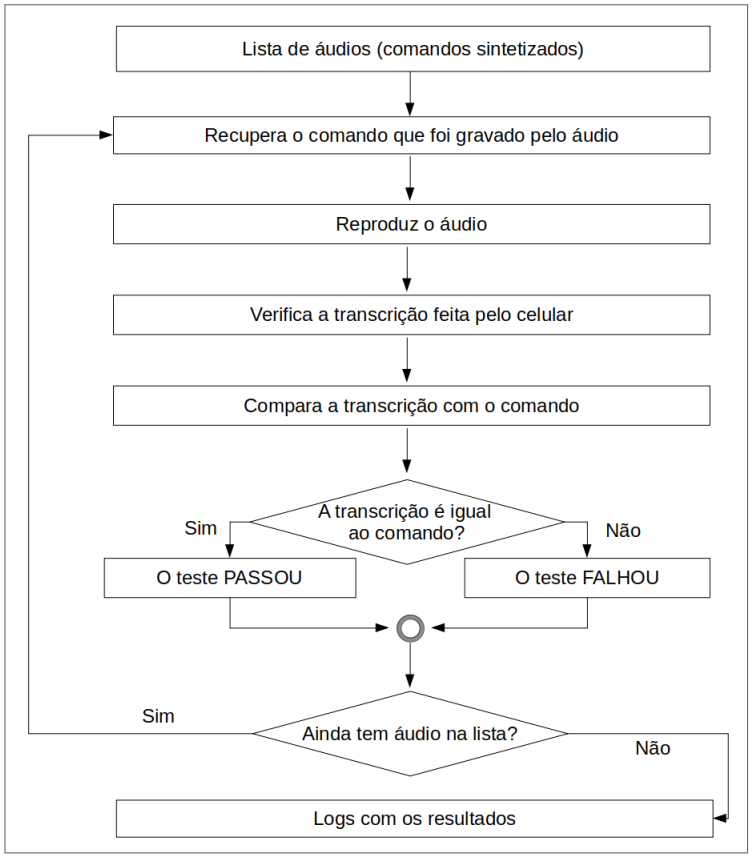


Figura 17 – Filtragem de Sentenças Equivalentes

Outro tipo de teste que também pode ser executado, é a verificação do retorno dado pelo sistema como um todo, isso já é feito na empresa. Dado um comando, é verificado se ele foi atendido pelo aparelho. Se tal comando for "ligar para Joanna", por exemplo, é verificado se a ligação foi feita depois do disparo dele. Tal tipo de teste é chamado de funcional.

4.4 CONSIDERAÇÕES FINAIS

Neste capítulo, foi apresentada a arquitetura para teste de sistemas *speech to text* resultante deste trabalho. Antes de tudo, foi dado o contexto geral, falando da problemática encontrada na execução de tais tipos de teste na empresa parceira e também, como poderia ser melhorada, o que deu origem à proposta. Com isso, foi apresentada a arquitetura desenvolvida, que foi dividida em cinco etapas principais, cujas foram detalhadas na seção de implementação.

5 EXPERIMENTOS E RESULTADOS

Os experimentos executados tiveram como objetivo principal verificar a clareza e naturalidade das falas sintetizadas. A finalidade era saber se os áudios sintetizados poderiam substituir os humanos nos testes de voz dos celulares da Motorola, sem impactar na qualidade de tais testes. Os experimentos feitos serão detalhados nas seções a seguir.

5.1 EXPERIMENTO I

Este foi o primeiro experimento a ser executado. Ele foi feito durante o início do desenvolvimento do módulo de sintetização da fala (Seção 4.3.4). Na época, a arquitetura proposta suportava apenas a síntese de falas do Amazon Polly.

O objetivo foi verificar a inteligibilidade (compreensão) por um sistema STT, das falas sintetizadas, afim de ter uma pré-validação do uso deste tipo de abordagem nos testes de voz. Também foi calculada a correlação entre os resultados dos testes executados com falas sintéticas e humanas, analisando, assim, se testes executados com dados sintetizados teriam resultados similares a testes habituais, com áudios gravados por humanos. O sistema de reconhecimento utilizado neste experimento foi o de um *smartphone* da Motorola que é incorporado a um aplicativo chamado de Moto Voice.

O Moto Voice é responsável por receber determinada requisição do usuário via voz, trascrever, interpretar e executar a tarefa relacionada. Antes mesmo do usuário falar aquilo que ele deseja que seja executado, ele precisa "desbloquear"tal aplicativo. Para isso, ele usa uma frase que foi previamente gravada que vai servir como chave de desbloqueio. Depois que tal frase é falada, uma tela de escuta é aberta pronta para receber o comando a ser executado. Tal comando pode ser alguma pergunta sobre a previsão do tempo, ou até mesmo pra abrir outro aplicativo, ou ainda para fazer alguma configuração no celular. Num cenário ideal, depois que o comando (de voz) é trascrito (para texto) e interpretado, a tarefa referente a ele é executada.

5.1.1 Protocolo Experimental

Para executar este experimento, foram sintetizadas os 13 comandos (sentenças) mais frequentemente usados nos testes de voz em celulares da empresa em questão. Arquivos de áudio foram gerados para tais comandos utilizando as 16 vozes dos diferentes tipos de inglês do Amazon Polly (Tabela 1).

Os testes foram executados manualmente em uma sala com pouco ruído, chamada *quiet room*, onde são feitos testes reais na Motorola. Foi utilizado um computador com uma caixa de som conectada, um celular e dois tripés, um para o celular e outro para a caixa de som, que distanciavam $\approx$ 1,85m entre si.

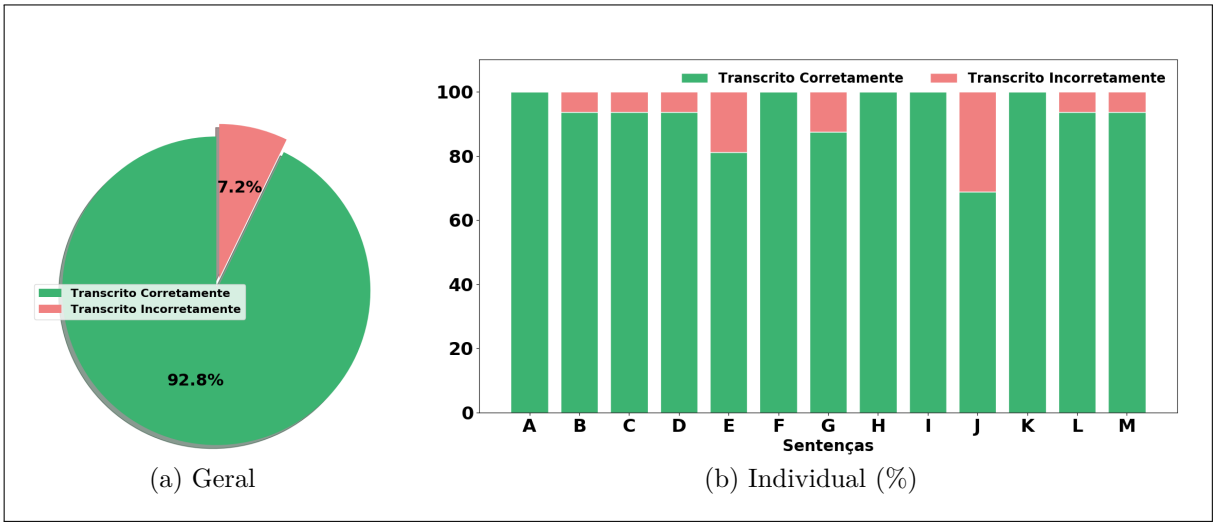
Durante a execução dos testes, para cada áudio reproduzido, era verificado se a tradução da fala para texto, pelo sistema do celular, estava correta e também, se o retorno do comando de voz era conforme o esperado. Ou seja, além da transcrição, era observado se o sistema respondia corretamente ao comando que estava no áudio, o que é denominado de teste funcional. Era caracterizada uma falha se o sistema não retornasse o resultado esperado para o respectivo comando, ou seja, se ele não funcionasse de acordo com as especificações.

Dos 13 comandos testados com falas sintetizadas, 7 foram executados por 12 humanos. Tal teste foi executado no mesmo ambiente (a *quiet room*), com o celular com as mesmas configurações, versão do sistema e dos aplicativos (que foi utilizado nos testes com uso de falas sintéticas). Da mesma maneira, era considerada uma falha se esse resultado fosse diferente do esperado.

Não foi possível executar todos devido ao gasto de tempo, uma vez que, só funcionários poderiam contribuir devido a confidencialidade e isso demandava tempo deles. Além disso, tinha a questão do uso do ambiente, a *quiet room*, que era dividida com outros times que fazem execução de teste de voz e dependendo do dia, a demanda é alta.

5.1.2 Resultados

Os resultados deste experimento responderam de forma positiva sobre a inteligibilidade dos áudios sintetizados. O total de áudios referentes aos comandos sintéticos testados foi de 208 (13 comandos × 16 locutores). Levando em consideração esse total, a figura 18a mostra que 92,8% (193 de 208) desses comandos foram corretamente entendidos pelo sistema STT, enquanto o restante, não (15 de 208). Enquanto na 18b, é possível ver a quantidade de comandos corretamente e incorretamente transcritos por sentença. Os comandos A, F, H, I e K foram reconhecidos de forma correta 100% das vezes.



(a) Geral (b) Individual (%)

Figura 18 – Porcentagem de sentenças corretamente transcritas

Em contrapartida, o comando J foi o que teve maior porcentagem de transcrições incorretas. Tal falha também ocorria quando o teste era feito com a voz humana. O motivo era que o sistema estava errando na transcrição da primeira palavra da sentença. Ele estava reconhecendo outra palavra, tanto quando recebia as falas sintéticas, quanto as humanas. Se os áudios que foram reconhecidos incorretamente referentes a tal sentença não forem levados em consideração (5 de 16), a porcentagem de reconhecimentos corretos sobe para 95%.

Quando é em levado em consideração o número de transcrições corretas e incorretas por locutor, observa-se que o sistema teve mais dificuldade de entender a fala do Ivy (Figura 19). Mais de 50% dos comandos falados por ele não foram transcritos corretamente, isso corresponde a 7 de 13 comandos.

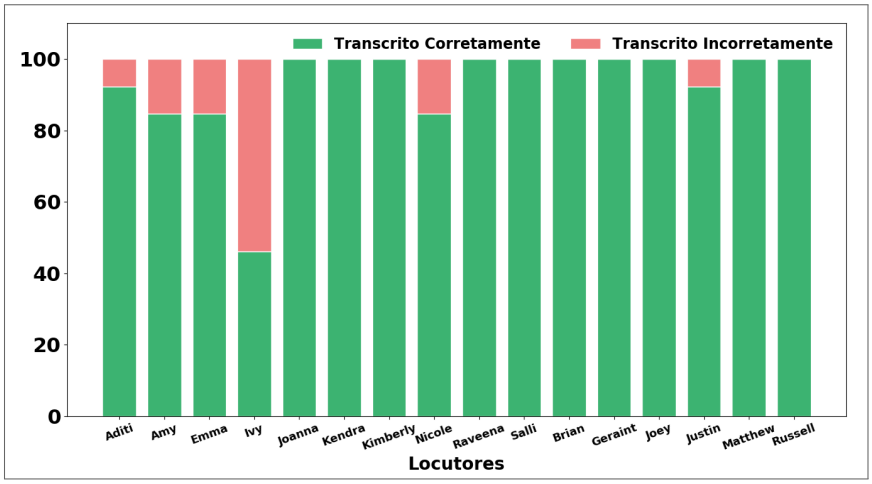


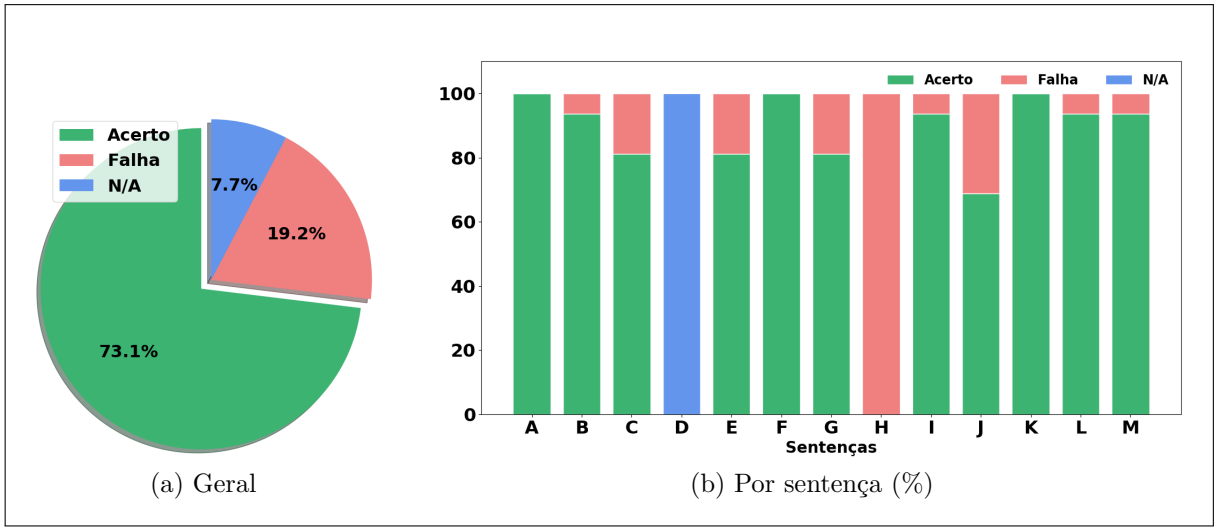
Figura 19 – Porcentagem de transcrições corretas e incorretas por locutor

Ainda na Figura 19, pode-se ver que a maior parte dos comandos não entendidos corretamente são de vozes do sexo feminino, que abrange o intervalo de Aditi a Salli. Isso corresponde a 14 daqueles 15 comandos. Apenas 1 é referente a uma fala masculina, a do Justin.

Além da transcrição das falas, foi verificada ainda a quantidade de erros funcionais por comando (ver Figura 20). Na Figura 20b, é possível ver que os testes dos comandos A, F e K tiveram 100% de acerto. Do total de corretamente reconhecidos, o, B, L e M tiveram também 100% de acerto no teste funcional, pode-se ver isso observando que a mesma porcentagem de comandos incorretamente transcritos Figura 18b, é igual a porcentagem de falhas para cada um (Figura 20b).

O comando D, embora tenha sido corretamente reconhecido quase 100%, das vezes, não foi possível verificar o resultado dado por ele por limitação de área. Só era possível verificar isso se a execução tivesse sendo feita nos Estados Unidos. O J foi o que teve maior porcentagem de erro de transcrição e foi o segundo que mais falhou de forma geral, na resposta dada pelo celular, a verificação funcional. Como o comando era traduzido de forma incorreta, o retorno que o aparelho dava também era incorreto, Figura 20b. O H

foi o que mais falhou, 100% dos testes referentes a tal sentença falharam. Foi conversado com um engenheiro de teste de voz da empresa sobre tal falha e foi dada informação de que realmente o sistema estava com bug relacionado a tal comando e que ele tinha sido reportado pouco tempo antes de tal conversa.



(a) Geral (b) Por sentença (%)

Figura 20 – Porcentagem de Sucesso e Falha

Sobre as falhas referentes aos testes funcionais com falas sintetizadas (19%, Figura 20a), foi observado que 35% delas foi devido à transcrição incorreta (Figura 21) e as demais, devido a bug do sistema, no caso da sentença H que falhou 100% das vezes; ou a outro tipo de falha que embora o sistema reconhecesse corretamente a sentença, às vezes ele fazia a ação desejada e outras, não. Nesse último caso, ele ficava parado ou dava uma resposta diferente da esperada.

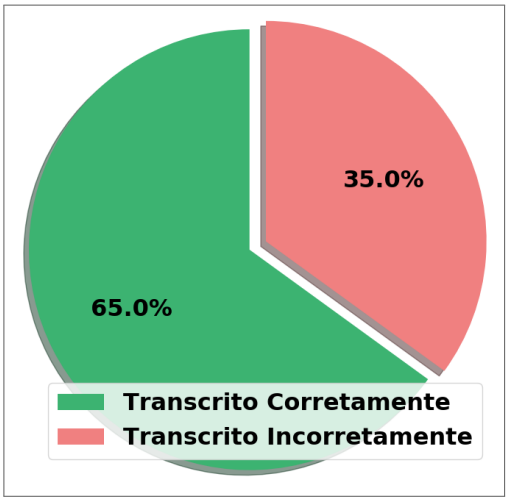


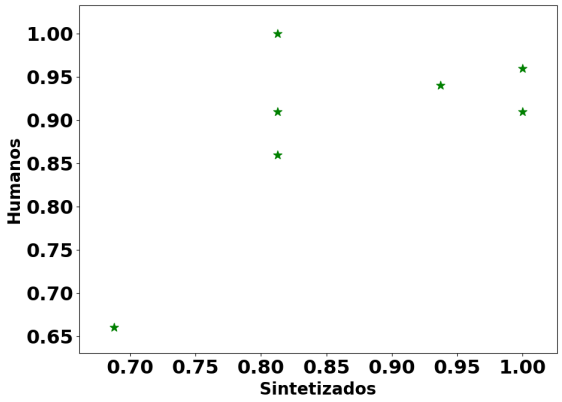
Figura 21 – Porcentagem de transcrições corretas e incorretas, levando em consideração apenas as falhas no teste funcional

Embora apenas 7 dos 13 comandos foram também executados com falas humanas, foi possível ver uma correlação favorável com os dados coletados das execuções de tais

testes com os feitos com as falas sintéticas. A Figura 22b mostra a correlação entre as 7 sentenças que foram executadas com humanos e também com falas sintetizadas. O eixo X representa a porcentagem de acerto dos casos de teste com falas sintéticas e o Y, com humanas. Cada estrela representa uma frase. Quanto mais próxima a estrela está da diagonal, maior a correlação dos dados. A correlação (de Pearson) calculada foi de 0,66, que segundo a escala, é moderada.

Sentença	Sintetizado	Humano
A	100,00%	91,00%
C	81,25%	86,00%
E	81,25%	91,00%
F	100,00%	96,00%
G	81,25%	100,00%
J	68,75%	66,00%
M	93,75%	94,00%

(a)



(b)

Figura 22 – Correlação entre as execuções dos testes funcionais com falas sintetizadas e humanas

Como o valor da correlação é positivo, têm-se que, à medida que a porcentagem de acerto relacionado a determinado comando com falas humanas cresce, a com falas sintetizadas também cresce. Da mesma forma, se um desses valores diminuir, também acontecerá o mesmo com o outro. Tais valores movem-se juntos (crescem ou decrescem) em uma proporção moderada.

Estes foram os resultados do primeiro experimento que impulsionaram mais ainda a pesquisa neste contexto, de usar áudios sintetizados ao invés de humanos na execução dos testes de sistemas STT. Assim, foram feitas outras avaliações usando outras abordagens e mais três serviços de sintetização que serão apresentadas na Seção 5.2.

5.2 EXPERIMENTO II

Tendo os resultados do experimento I e vendo a possibilidade de se usar falas sintetizadas nos testes de voz, surgiu a necessidade de realizar mais experimentos, com mais dados e outras abordagens de avaliação. E ainda, com a arquitetura mais completa, robusta e com mais opções de serviço para sintetização. Assim, este experimento tem o objetivo de verificar a utilidade de áudios sintéticos como alternativa para os gravados nos testes de sistemas STT.

Para tal, foram feitas análises objetivas e subjetivas. As objetivas contaram com a participação de usuários para julgar os áudios, tanto sintéticos quanto humanos. Já as subjetivas dizem respeito à execução de testes automáticos no celular utilizando os dois tipos de áudio. Detalhes de cada avaliação serão vistos nas Seções 5.2.1 e 5.2.2.

5.2.1 Avaliação Subjetiva

Esta avaliação foi baseada na opinião das pessoas a respeito dos áudios. Assim, foi solicitado que colaboradores da empresa parceira para fazê-la. Para isso, foi implementado um sistema onde o usuário pudesse interagir e dar suas respostas. Ao todo, 50 pessoas se disponibilizaram para participar do experimento. Os resultados coletados, bem como o protocolo experimental, serão apresentados ao decorrer desta seção.

5.2.1.1 Protocolo Experimental

A avaliação subjetiva contou com duas fases diferentes, uma referente ao Teste de Turing e outra, ao de Qualidade. No teste de Turing, o usuário ouvia cada áudio selecionado e respondia se era um humano ou uma máquina que estava falando. No de Qualidade, ele dava uma nota de 1 a 5 que correspondia à qualidade do áudio ouvido. Tal qualidade englobava naturalidade e inteligibilidade.

Organização das Bases de Áudios e Seleção Estratificada

Para o levantamento e organização da base de áudios, foram escolhidas 12 sentenças diferentes em inglês, de acordo com os comandos gravados que tinham no banco de dados da Motorola para teste de voz. Assim, foram coletados todos os áudios gravados disponíveis no banco, referentes a estes comandos e localidades que falam inglês. E, assim, a base de áudios humanos foi coletada.

Por outro lado, para a construção da base de falas sintéticas, foram gerados áudios referentes a aquelas 12 sentenças, usando as quatro abordagens de sintetização (A. Polly, Google TTS, IBM Watson TTS e M. Azure TTS), e todos os tipos de inglês disponíveis nos respectivos serviços.

Na Tabela 11, a quantidade de áudios disponíveis está detalhada, dividida por tipo e por serviço. Lembrando que a quantidade dos sintetizados depende da quantidade de vozes disponíveis em cada um. Como o da Watson é o que tem menos, apenas 4, foram gerados 48 áudios referentes às 12 frases. A mesma lógica é aplicada para os outros serviços.

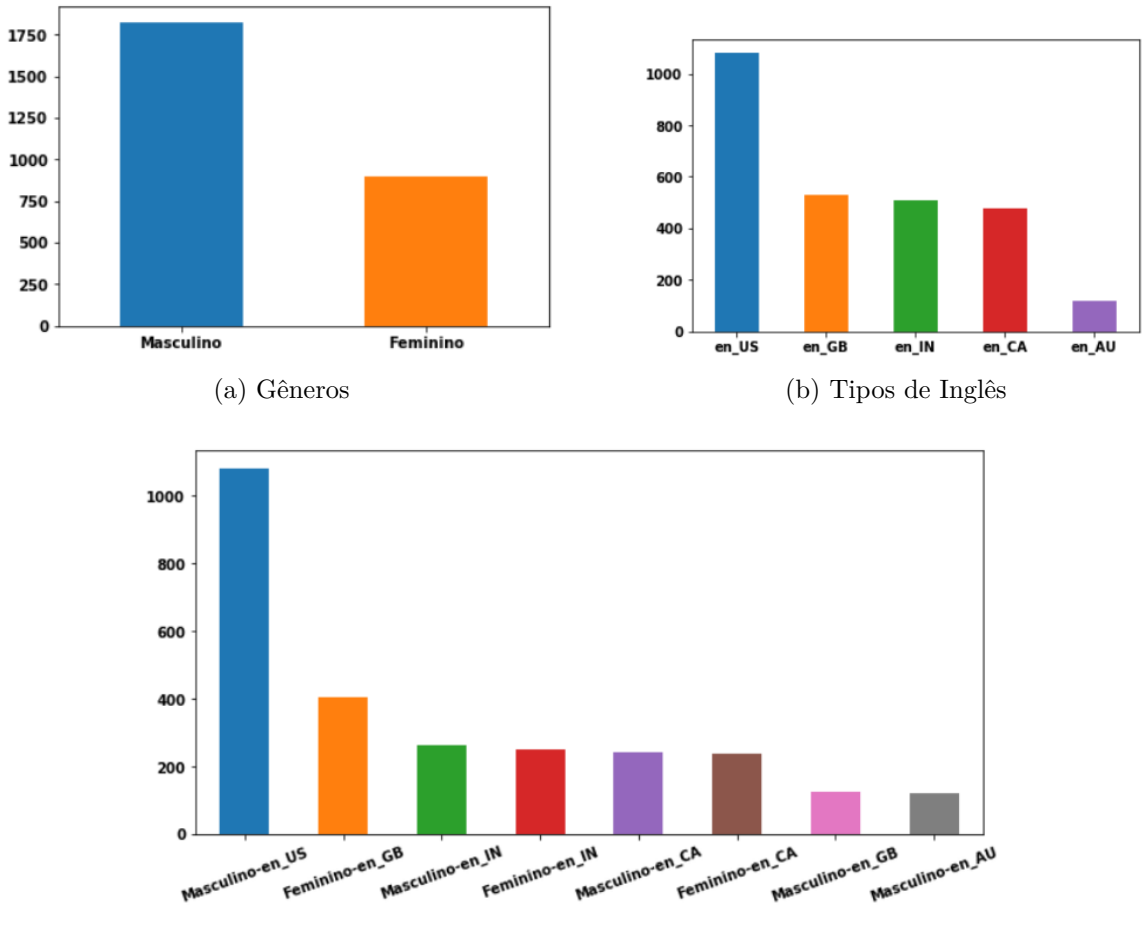
Como a quantidade de áudio é desbalanceada (existem mais áudios humanos que sintetizados; e diferentes quantidades para cada serviço) e era inviável para cada voluntário avaliar os 3462 áudios, foi aplicada seleção estratificada para a escolha dos áudios que cada pessoa iria julgar. Além disso, foi definido que 20 áudios seria uma quantidade adequada para um voluntário avaliar em cada fase, de forma que não fosse enfadonho e que ele não

perdesse o interesse de responder; e, por outro lado, ele poderia avaliar uma variedade de áudios de acordo com a seleção.

Tabela 11 – Quantidade de áudios humanos (gravados) e sintetizados utilizados

Gravados	**Sintetizados**				**Total**
	A. Polly	Google TTS	IBM Watson TTS	M. Azure	
	192	312	48	192	
2718	**744**				**3462**

Fonte: Este trabalho



(a) Gêneros

(b) Tipos de Inglês

(c) Classes

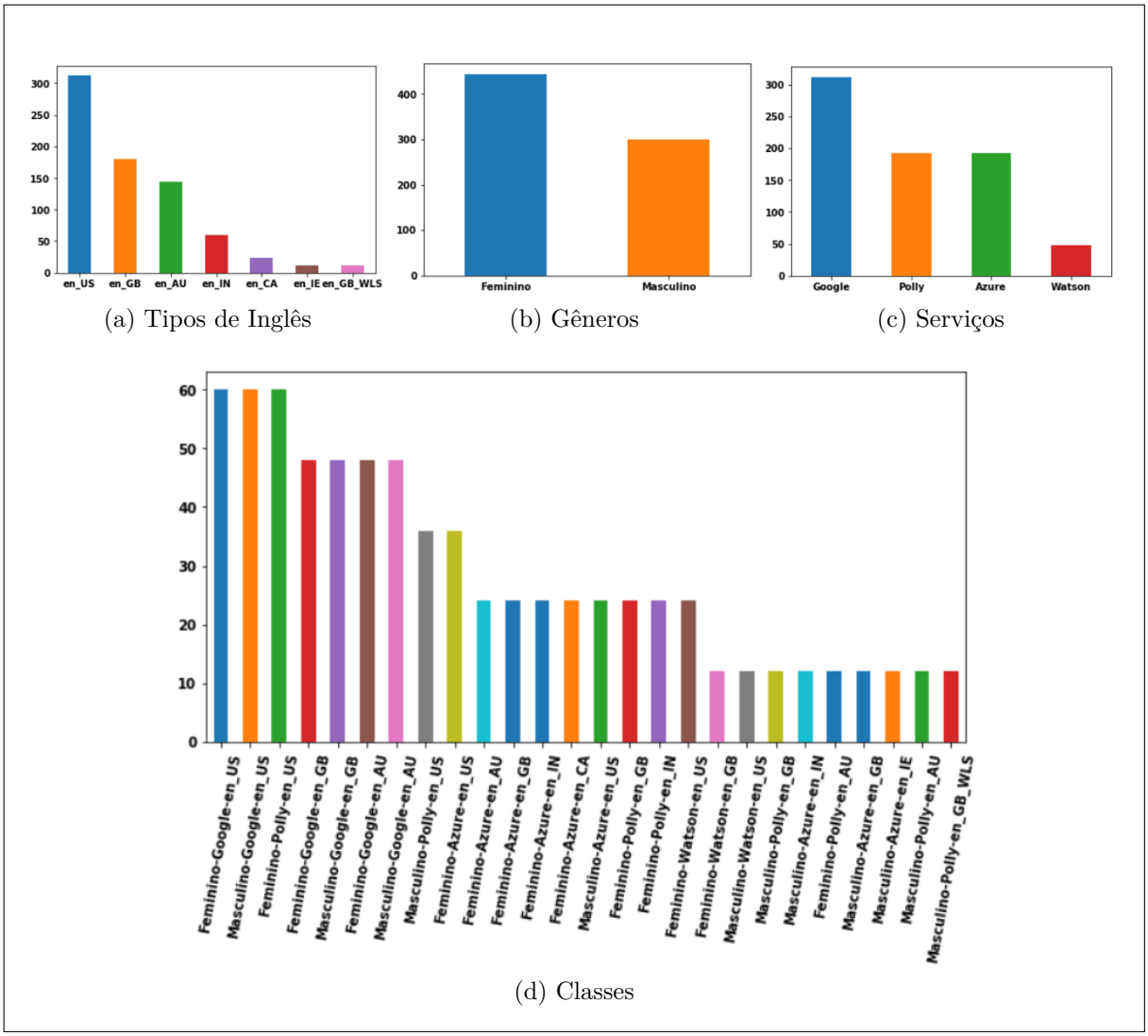
Figura 23 – Histogramas dos áudios gravados

Para que cada usuário pudesse avaliar áudios diversificados, sem correr o risco de todos os áudios escolhidos serem do inglês americano, por exemplo, ou de um só gênero, etc, para cada seleção, foi levada em consideração a variedade dos tipos de inglês e gênero dos áudios humanos. Já para os sintéticos, além dessas duas características, também foi

considerado o serviço de sintetização da fala. A partir disso, foram definidas algumas classes de áudios.

Na Figura 23 são apresentados três histogramas referentes aos gêneros, tipos de inglês e classes dos áudios humanos, respectivamente. Observa-se que dos 2718 áudios gravados, mais de 1750 são masculinos. Em relação à linguagem, mais de 1000 são referentes ao inglês americano, e o restante é dividido para os outros quatro tipos.

Para fazer a seleção estratificada, foram definidas as classes (Figura 23c), baseando-se no gênero e tipo inglês, usando a seguinte estrutura: gênero-inglês. A primeira classe "Masculino-en_US", por exemplo, representa os áudios masculinos do inglês americano, que, por usa vez, é a classe majoritária.



(a) Tipos de Inglês (b) Gêneros (c) Serviços

(d) Classes

Figura 24 – Histogramas dos áudios sintetizados

Para a formação das classes dos áudios sintéticos também foi levado em consideração o serviço de sintetização, Figura 24. Assim, a organização ficou da seguinte forma: gênero-serviço-inglês. A primeira classe mostrada na Figura 24d representa os áudios femininos

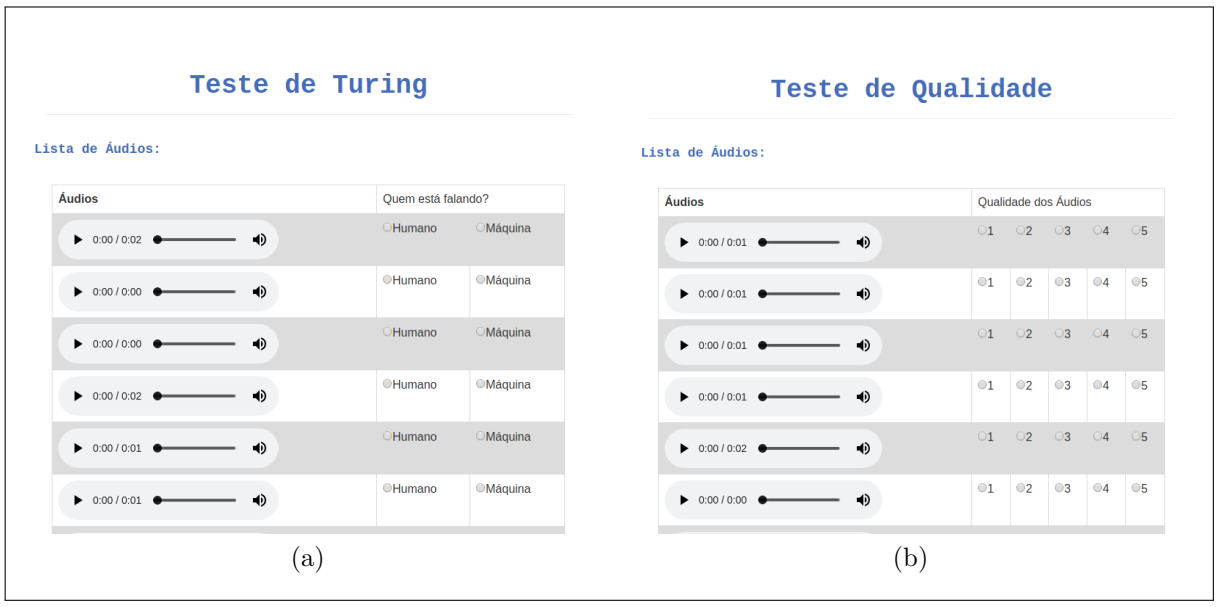
sintetizados usando o serviço da Google referentes ao inglês americano.

As classes definidas foram 34, 8 dos áudios gravados e 26 dos sintéticos, o que excedeu a quantidade (20) de áudios que cada voluntário escutaria em cada experimento. Tendo em vista que a porcentagem de classes humanas é quase 25%, esta foi a proporção de falas humanas que seria avaliada, no caso 75% dos áudios (*i.e.*, 15 áudios) apresentados para cada usuário seriam selecionados de classes aleatórias de áudios sintetizados enquanto que 25% dos áudios (*i.e.*, 5 áudios) seriam selecionados dentre classes aleatórias de áudios gravados.

Avaliação Feita pelos Voluntários

Cada usuário logava no sistema e preenchia algumas informações pessoais (*i.e.*: nacionalidade, nível de inglês, idade). Depois, era direcionado para uma tela que explicava sobre as fases do experimento. Em cada uma dessas duas etapas, 20 áudios eram escolhidos (para cada usuário), de forma estratificada, conforme foi explicado. Cada áudio poderia ser ouvido até três vezes, depois disso, ele era desativado.

A primeira fase era o teste de Turing. Nela, foi perguntado se era um humano ou uma máquina que estava falando em cada áudio. Assim, havia o conjunto de áudios e ao lado de cada um, duas opções de resposta (humano e máquina) (ver Figura 25a). Após ouvir o áudio, o voluntário dava a resposta.



(a) (b)

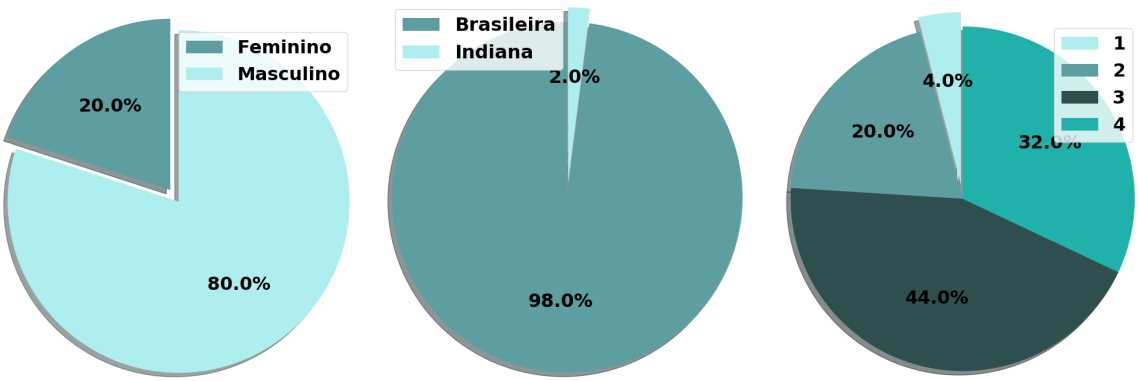
Figura 25 – Correlação entre as execuções dos testes funcionais com falas sintetizadas e humanas

Terminando o teste de Turing, ele partia pra próxima, a de Qualidade. Da mesma forma que na primeira etapa, havia um conjunto de áudios, mas o que foi perguntando foi sobre a qualidade (naturalidade e inteligibilidade) de cada um deles. Assim, o avaliador

poderia dar uma pontuação de 1 a 5 (Figura 25b), onde 1 indicava um áudio com qualidade péssima e 5, ótima. Finalizando esta etapa, ele era levado para uma tela de feedback, onde poderia falar sobre sua experiência. Depois, a avaliação era finalizada. Uma vez feita, ela não poderia ser feita novamente pelo mesmo usuário. Os resultados desta avaliação serão apresentados e discutidos a seguir.

Características dos Voluntários

As pessoas que participaram dos testes de Turing e de qualidade, avaliando os áudios, foram funcionários da empresa parceira, Motorola. Ao todo, cinquenta pessoas participaram, sendo a maioria do sexo masculino, 80% e 98% de nacionalidade brasileira, conforme pode ser visto nas Figuras 26a e 26b, respectivamente.



(a) Gênero (b) Nacionalidade (c) Nível de Inglês

Figura 26 – Características dos voluntários que avaliaram os áudios

Como os áudios avaliados eram em inglês, algo a se considerar era se os voluntários sabiam tal língua. Embora isso seja um pré-requisito para eles, pois têm que estar diariamente se comunicando com pessoas de outros países usando tal língua como meio, muitas vezes tal comunicação é por meio da escrita, não necessariamente falada (ouvida), assim era importante saber como eles se nivelavam.

Os níveis de inglês perguntados e informados variaram de um 1 a 4 (Figura 26c), onde o número 1 representa o nível básico e 4, fluente. Os níveis 2 e 3 são o intermediário e avançado, respectivamente. A maioria das pessoas, 44%, se declararam ser do nível avançado, enquanto 32% falaram que são fluentes. Apenas 4% consideram seu nível de inglês básico, enquanto os 20% restantes, se disseram intermediários.

5.2.1.2 Resultados

Esta subseção foi dividida em duas partes, duas referentes aos dois tipos de avaliação feitas pelos voluntários, o teste de Turing e de Qualidade, onde, em cada uma destas

partes serão apresentados gráficos correspondentes aos resultados obtidos, bem como a discussão dos mesmos.

Teste de Turing

Neste teste, cada voluntário escutou 20 áudios e julgou cada um deles como sendo uma fala humana ou sintetizada. Considerando o processo de sintetização, é considerado sucesso quando um áudio sintetizado é confundido como sendo gravado por um humano. Taxa de sucesso de uma ferramenta de sintetização será então o percentual de vezes em que os áudios sintetizados apresentados para os usuários foram julgados como sendo áudios gravados por humanos. Como base de comparação, verificamos também o percentual de áudios humanos apresentados para os usuários, que foram confundidos como sendo gerados por sintetização (Tabela 12).

Tabela 12 – Significado das labels de falha e sucesso

Label	**Origem do áudio**	**Resposta dada**
Falha	Gravado	Gravado
	Sintetizado	Sintetizado
Sucesso	Sintetizado	Gravado
	Gravado	Sintetizado

Fonte: Este trabalho

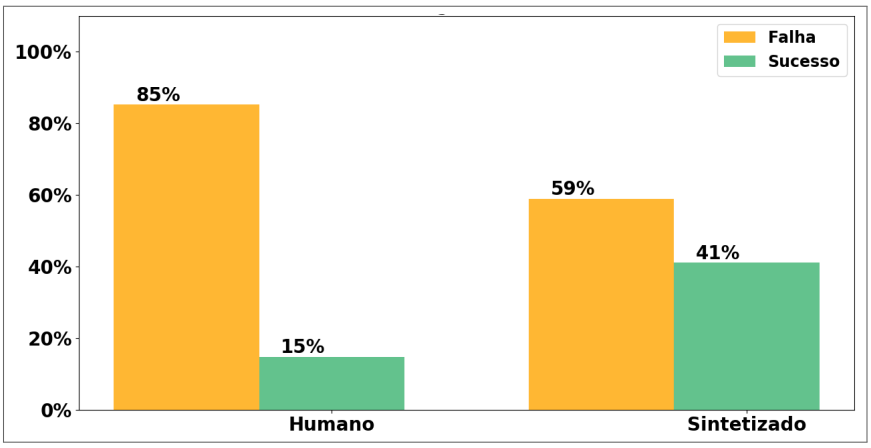


Figura 27 – Taxa de falha e sucesso geral e por tipo (humano e sintético)

A Figura 27 apresenta um gráfico com as taxas de sucesso geral, de todos os áudios analisados no teste de Turing, tanto os humanos, quanto os sintetizados. As porcentagens estão divididas por tipo de áudio (humanos e sintetizados). Houve 15% de sucesso dos áudios humanos analisados, como esperado. Já nos sintéticos, tal porcentagem foi de

59%. Assim pode-se concluir que os áudios humanos são relativamente fáceis de serem identificados, que naturalidade é um fator consideravelmente importante e que pesou no momento da avaliação.

Na Figura 28, são apresentados os resultados dos áudios sintéticos por gênero masculino e feminino. A taxa de sucesso dos áudios femininos foi de 30%, valor menor que a dos masculinos, que foi de 54%. Isso indica que as pessoas mais julgaram errado do que certo os áudios com falas masculinas. Eles causam mais dúvida no geral, conseguem confundir bem.

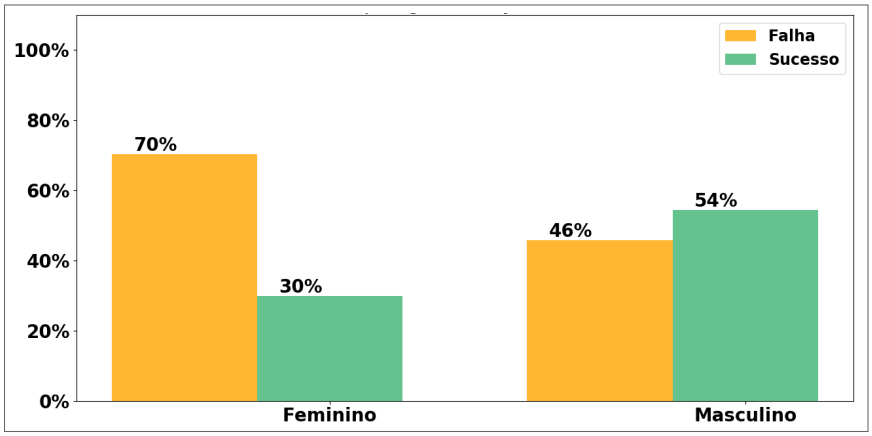


Figura 28 – Taxa de falha e sucesso por gênero dos áudios sintéticos

Analisando os áudios sintéticos divididos por serviço (Figura 29), observa-se que os da Azure são os que tiveram a menor porcentagem de sucesso no geral (32%). Já os da Google foram os que tiveram maior, tendo tal taxa superado a de falha, 53% e 47%, respectivamente. Isso indica que as falas sintetizadas por tal serviço conseguem ser confundidas bem com as humanas. Os da Polly e Watson foram os que tiveram taxas de engano intermediárias, totalizando 42%.

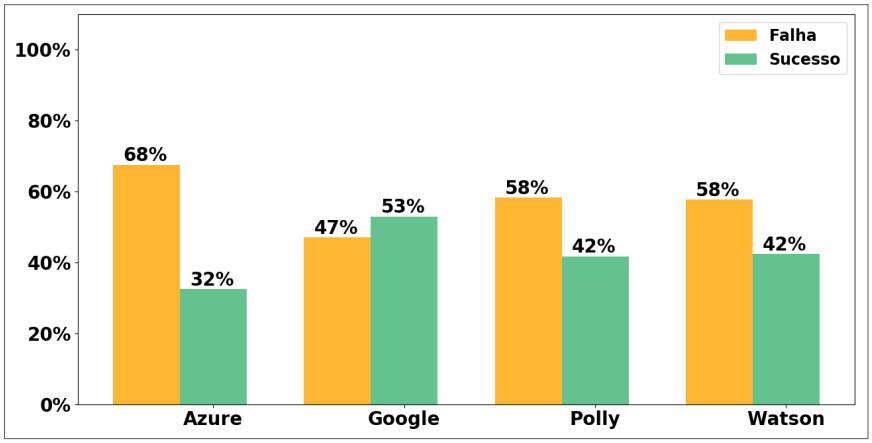
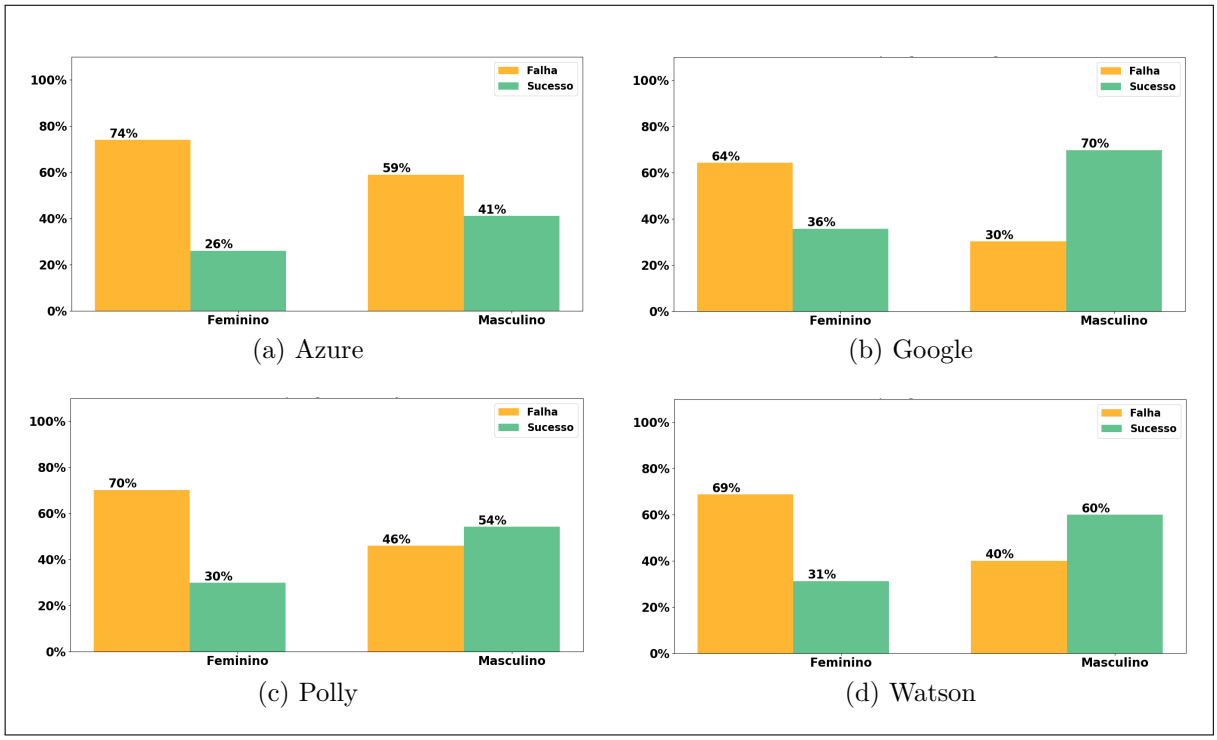


Figura 29 – Taxa de falha e sucesso por serviço

Na Figura 30 estão detalhadas as porcentagens de sucesso e falha referentes às falas de cada serviço por gênero. As taxas de sucesso relacionadas às vozes femininas em todos os

serviços são menores que as de falha, confirmando, mais uma vez que as falas sintetizadas de tal gênero não são tão similares às humanas. De todas as taxas de sucesso relacionadas às vozes de tal gênero, a do serviço Azure é a porcentagem menor e a da Google, maior.

Em contrapartida, as porcentagens de sucesso relacionadas às vozes masculinas são maiores que as de falha em todos os serviços, exceto o da Azure, com 41% de sucesso. Tal taxa referente às da Polly e Watson totaliza 54% e 60%, respectivamente. Já em relação aos da Google, essa porcentagem chega a atingir 70%, a taxa mais alta de sucesso indicada até o momento.



(a) Azure (b) Google

(c) Polly (d) Watson

Figura 30 – Taxa de falha e sucesso por serviço por gênero

Conforme já mencionado, com o serviço da Google é possível sintetizar falas de dois tipos que são chamados de Padrão e Wavenet. A empresa alega que esse segundo tipo gera falas de maior qualidade. O que pode ser observado nos resultados do teste de Turing, Figura 31, é que as do tipo Wavenet são mais confundidas com falas humanas do que as Padrão, 59% e 46% de sucesso, respectivamente.

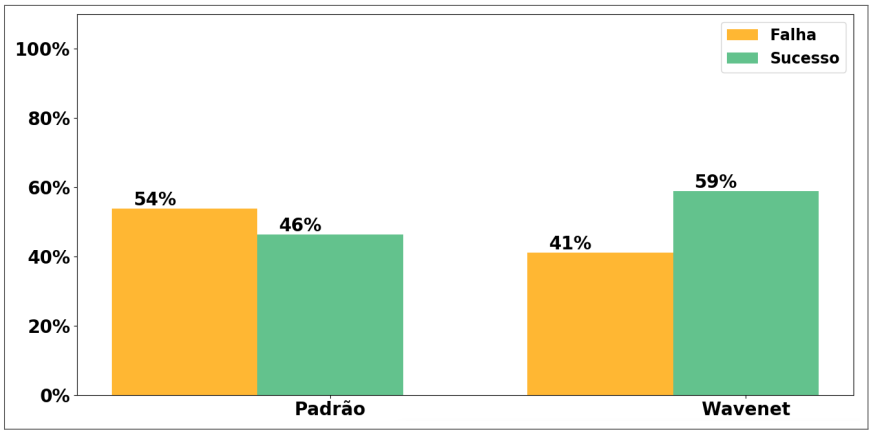


Figura 31 – Taxa de falha e sucesso (Google Padrão e Wavenet)

Na Figura 32 estão as taxas de sucesso e falha relacionadas às falas dos locutores da Azure. A taxa de sucesso referente à voz da primeira locutora (ZiraRUS) foi de 0%, ou seja, todas as avaliações feitas com a fala dela foram julgadas como sintética. Das 11 vozes femininas, em apenas um caso, a porcentagem de sucesso foi maior que a de falha (HayleyRUS). Em relação às 5 vozes masculinas, nenhuma teve 0% de sucesso. A menor porcentagem de sucesso foi de 28%, referente à fala de Ravi e duas delas tiveram tal porcentagem superior à de falha, as duas últimas com taxas 56% e 75%, respectivamente.

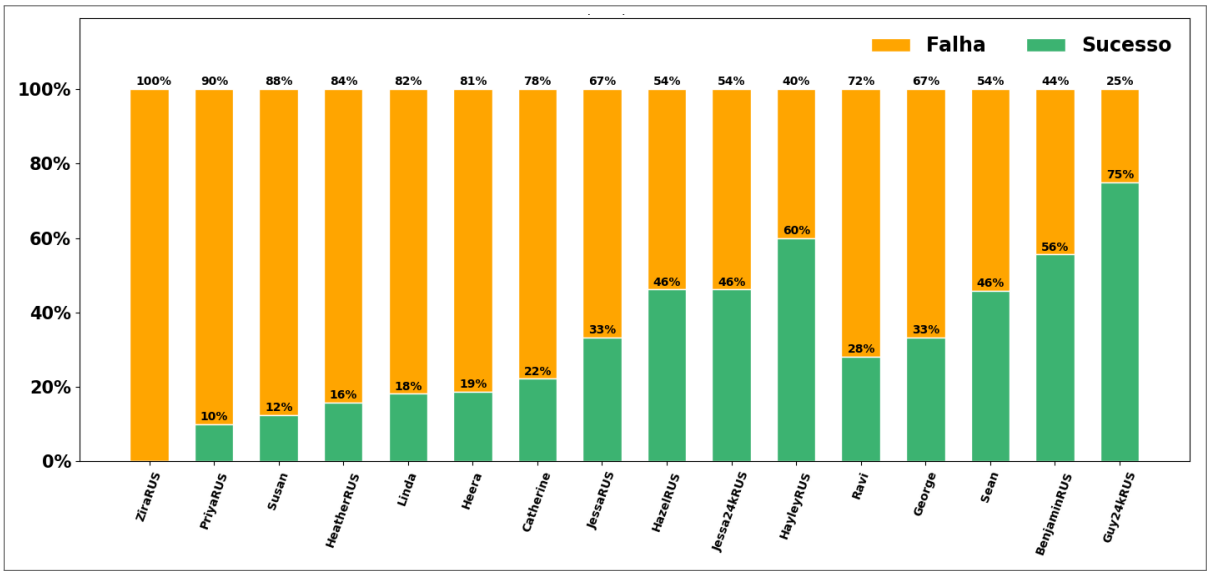


Figura 32 – Taxa de falha e sucesso por locutor - Azure

Na Figura 33, estão os resultados referentes às vozes da Google. Das 13 vozes femininas, 10 tiveram taxa de sucesso abaixo de 50%. Uma dessas teve 100% de falha, ou seja, foi identificada corretamente como voz sintética todas as vezes que foi julgada. E, apenas 3 tiveram taxa de sucesso maior ou igual a 50%.

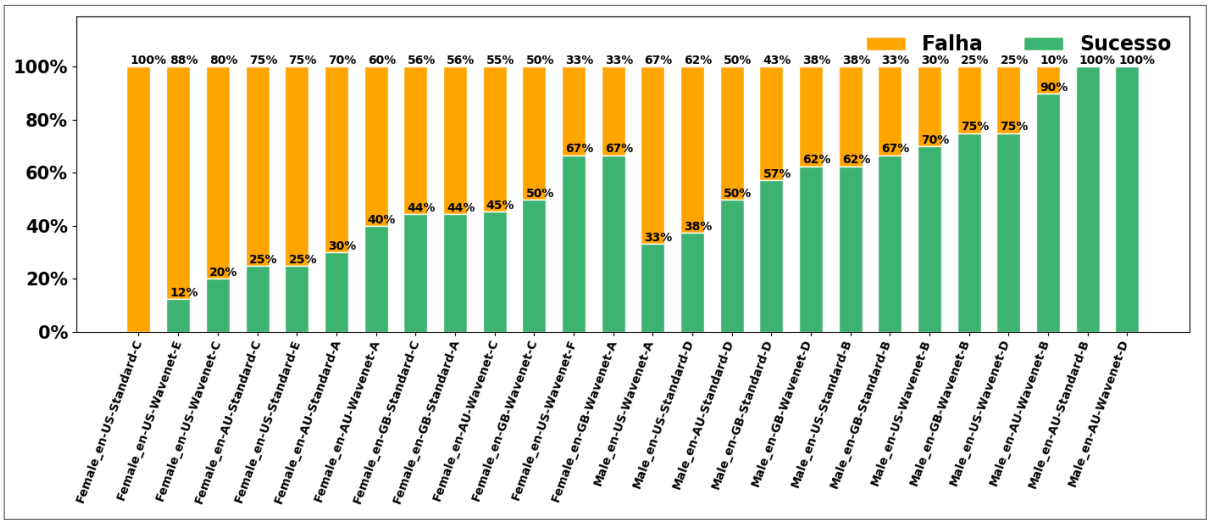


Figura 33 – Taxa de falha e sucesso por locutor - Google

Já das 13 vozes masculinas da Google, apenas 2 tiveram taxa de sucesso inferior à de falha. Mesmo assim, a menor porcentagem de sucesso foi de 33%, enquanto no outro gênero, tiveram 6 vozes abaixo desse percentual. As outras 11 vozes masculinas tiveram taxa de sucesso maior ou igual a 50%, sendo que, duas delas atingiram os 100%.

O gráfico referente aos locutores da Polly estão na Figura 34. Novamente, a taxa de sucesso referente a uma das vozes femininas foi 0%. Além disso, nenhuma da 10 vozes femininas de tal serviço teve taxa de sucesso acima de 50%. Das 6 masculinas, três tiveram taxa de sucesso abaixo da de falha e três, acima. Diferentemente do serviço da Google, nenhuma das vozes confundiu 100% das vezes em que foi avaliada. A maior taxa de sucesso foi de 67%.

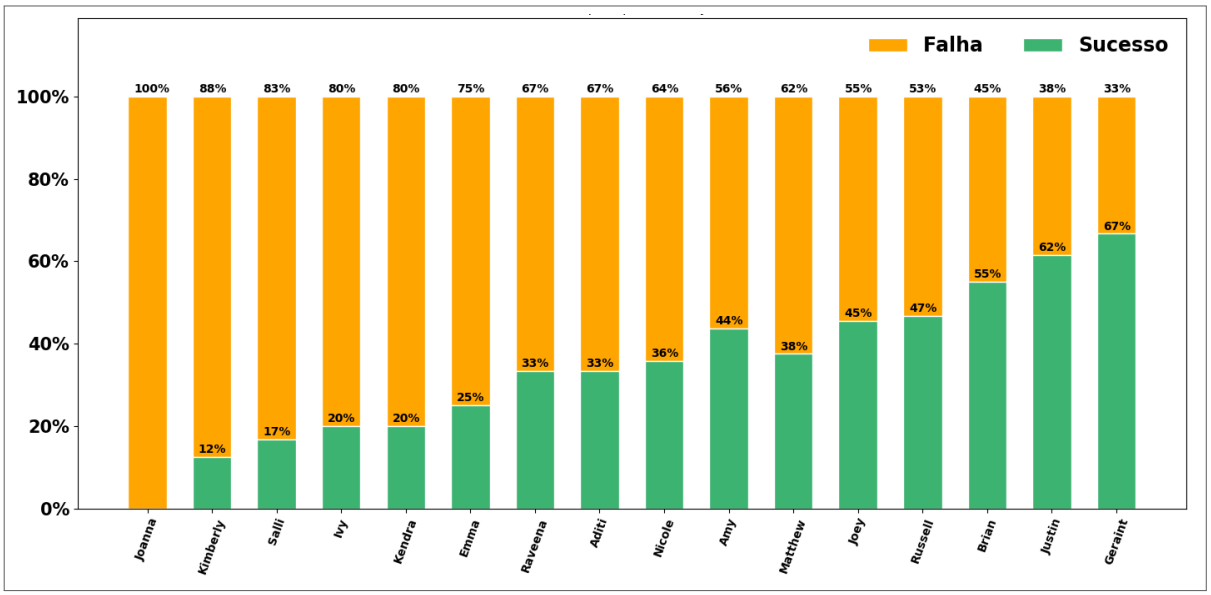


Figura 34 – Taxa de falha e sucesso por locutor - Polly

Na Figura 35 estão os quatro locutores da Watson e suas respectivas porcentagens de

falha e sucesso. São 3 vozes femininas e destas, apenas uma teve taxa de sucesso maior que a de falha, 46% e 54%, respectivamente. A única masculina também essa configuração, mas com taxas de 60% e 40% de sucesso e falha, respectivamente.

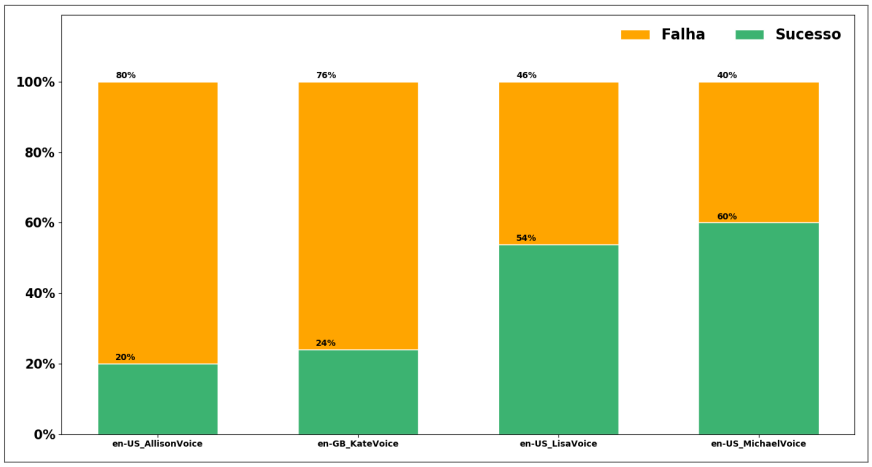


Figura 35 – Taxa de falha e sucesso por locutor - Watson

Foi feito também o gráfico de comparação entre as taxas de sucesso e falha por tipo (Padrão e WaveNet) de voz da Google (Figura 36). Das 12 vozes padrão, 7 tiveram taxa de falha superior à de sucesso. Enquanto das 14 Wanenet, apenas 5 tiveram tal resultado. Tal gráfico ainda mostra outra coisa. Observe que as duas piores vozes de cada tipo são femininas. Enquanto as duas melhores são masculinas.

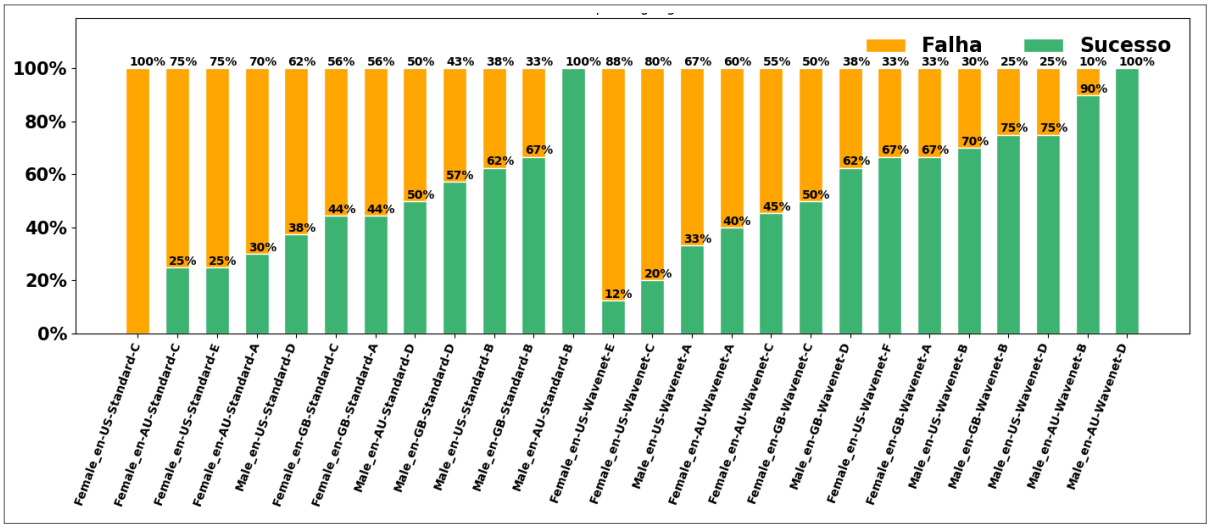


Figura 36 – Taxa de falha e sucesso por locutor - Google Padrão e Wavenet

O inglês que teve maior porcentagem de falha foi o Canadense (en-CA), com 83% (Figura 37). Já o que teve menor, foi o Galês (en-GB-WLS), com 33%. No entanto deve-se levar em consideração que a representatividade de tal inglês na base é baixa, só tem uma voz, a de Geraint. Tal voz, por sua vez, foi a que teve melhor taxa de sucesso de todas do serviço Polly (ver na Figura 34).

Outro inglês com poucas amostras é o da Irlanda (en-IE), com apenas uma voz que é da Azure, a de Sean (masculina). Já o Americano (en-US), Australiano (en-AU), Britânico (en-GB) são os que têm maior representatividade na base. Destes o com maior taxa de sucesso foi o en-AU.

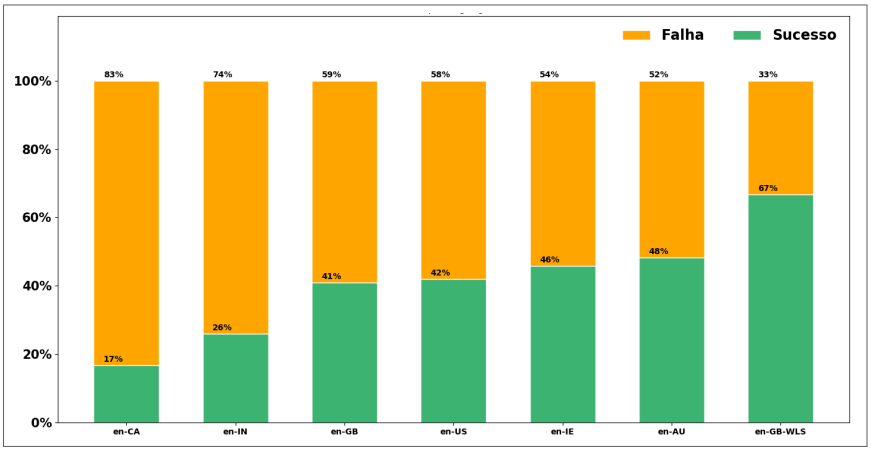


Figura 37 – Taxa de falha e sucesso por linguagem de todos os áudios sintéticos

Na Figura 38 está a média das respostas corretas dada pelos voluntários de cada nível de fluência. Observa-se que a maior porcentagem de respostas corretas foram das pessoas do nível básico, no entanto, apenas 4% dos voluntários declararam ter este nível, 1. Esta quantidade é pequena para se tirar uma conclusão mais concreta.

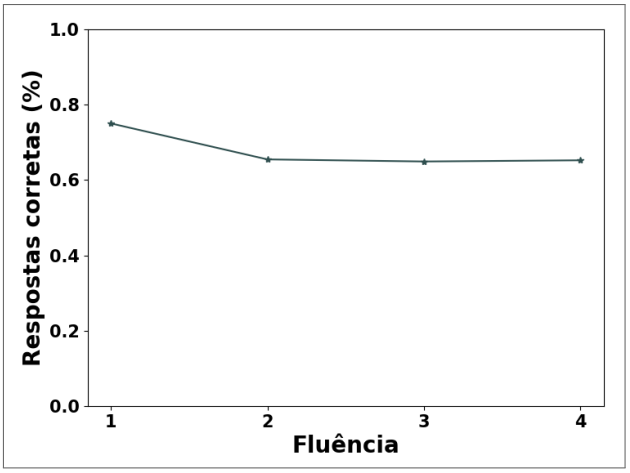


Figura 38 – Fluência dos voluntários x Respostas corretas no teste de Turing

Já a porcentagem de respostas corretas das pessoas de níveis 2, 3 e 4 é quase a mesma, todos com $\approx 65\%$. Neste caso, o nível de inglês que foi declarado por cada voluntário não impactou de forma visível a avaliação dada por eles.

O teste de Turing mostrou que os áudios femininos sintetizados, no geral, são relativamente distintos dos humanos, ou seja, as pessoas geralmente acertam na avaliação. Elas sabem dizer se uma fala sintetizada de tal gênero é humana ou não. Já os masculinos

causam mais confusão, ainda mais se forem geradas pelo serviço da Google e do tipo Wavenet.

Teste de Qualidade

Nesta fase, o voluntário tinha que escutar 20 áudios e dar uma qualidade a cada um deles numa escala de 1 a 5, onde 1 indica uma qualidade péssima e 5, ótima. A Figura 39 mostra a porcentagem de áudios com qualidade geral de cada nível. As maiores porcentagens de áudios humanos estão distribuídas nos níveis 2, 3 e 4 de qualidade, sendo que a maior é 29% e tem nível 3. Já a maior parte dos sintetizados têm a qualidade 4 e 5, englobando 32% e 39% dos áudios, respectivamente.

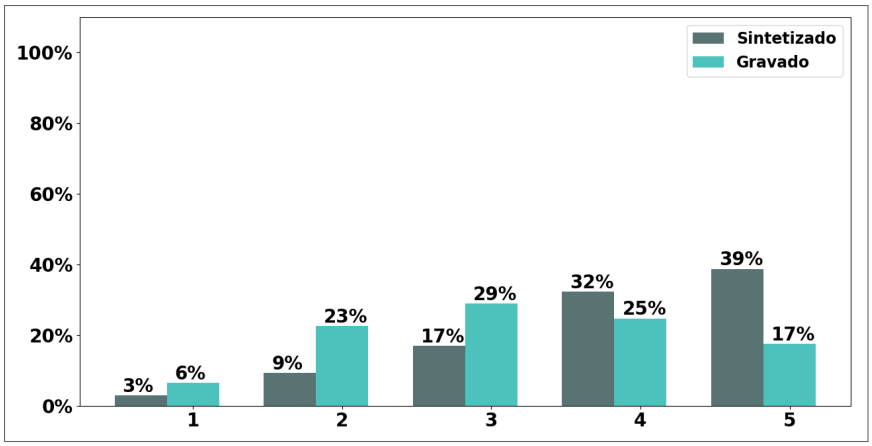


Figura 39 – Qualidade geral (humano e sintetizado)

Existem mais áudios femininos do que masculinos com qualidade 1, 2, 3 e 4, no entanto a diferença percentual não é muito grande, a maior é de 3%. Já com qualidade 5, existem mais áudios masculinos, 38%, enquanto 29% são femininos nesse nível (Figura 40).

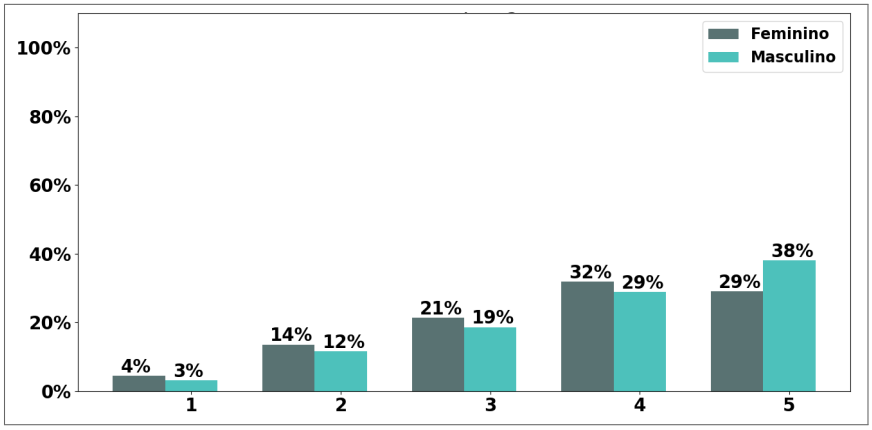


Figura 40 – Qualidade por gênero

Dos quatro serviços de sintetização, os da Google é o que tem mais qualidade. 86% de seus áudios estão concentrados nos níveis 4 e 5, sendo 28% no primeiro e 58% no segundo.

Os da Azure e Watson estão distribuídos entre os três mais altos: 3, 4 e 5. Já os da Polly, também estão concentrados entres os dois últimos

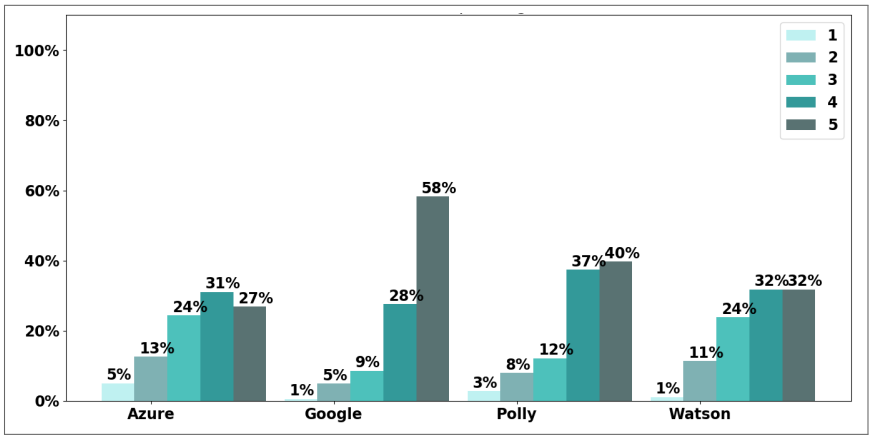


Figura 41 – Qualidade por serviço

Foi feita uma análise (Figura 42) da relação entre taxa de sucesso do teste de Turing (engano do homem pela máquina) e cada nível de qualidade dos áudios. Observa-se que os áudios com qualidade 1 (a pior) foram os que menos conseguiram enganar os voluntários, eles têm a menor taxa de sucesso ($\approx 21\%$). Tal valor se distancia em mais de 20% dos de melhor qualidade (5), que têm taxa de sucesso de $\approx 49\%$.

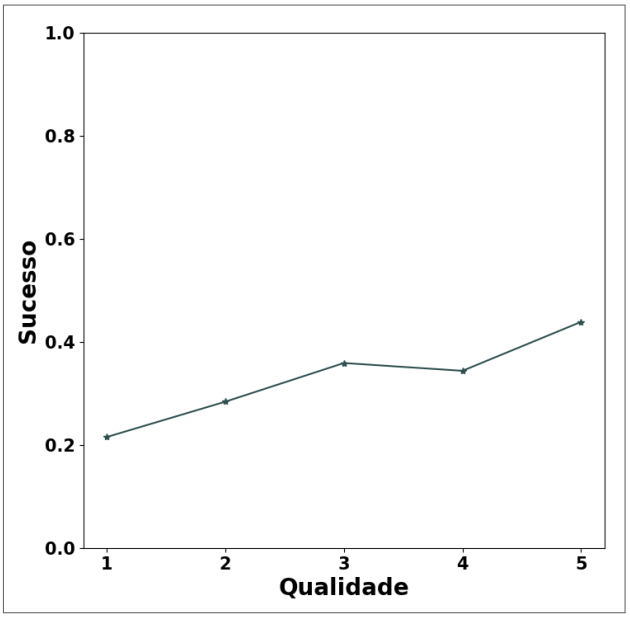


Figura 42 – Sucesso (Turing) x Qualidade

Ainda na Figura 42 é possível ver que as taxas de sucesso são ascendentes, conforme o aumento do nível de qualidade do 1 ao 3. Do 3 ao 4 tem uma queda de 36% para 34,4%, ou seja, não chega a 2% de diferença. E, novamente, do 4 ao 5, ocorre uma ascendência. Desta forma, tem-se que as pessoas conseguem identificar melhor a procedência do áudio quando ele tem a qualidade ruim.

5.2.2 Avaliação Objetiva

Na avaliação objetiva, foram feitos testes automáticos de dois tipos. O primeiro, visando verificar a transcrição e o segundo, a funcionalidade do sistema (a resposta dada a certo comando). Os resultados e o protocolo desta avaliação serão apresentados ao decorrer desta seção.

5.2.2.1 Protocolo Experimental

Os comandos testados foram os mesmos (12) da avaliação subjetiva (Seção 5.2.1), exceto um deles porque ele não estava sendo utilizado na versão do aplicativo (Moto Voice) em questão.

Tendo os scripts implementados e os áudios, os testes foram colocados para serem executados. O ambiente de execução foi caixa com isolamento acústico. Nela, foram colocados uma caixa de som e o celular com o sistema a ser testado, que estavam conectados a um computador localizado fora da caixa.

O primeiro tipo de teste a ser executado foi o de transcrição. Cada áudio foi reproduzido cinco vezes, retornando um status de falha ou sucesso em cada uma delas. O sistema falhava quando a transcrição feita do áudio era incorreta e era bem sucedido se ela fosse correta. Neste teste não era verificada a resposta dada pelo aparelho.

O segundo teste executado foi o funcional. Da mesma forma, cada áudio foi reproduzido cinco vezes e era verificado se ele passava ou falhava. Diferentemente do teste de transcrição, este verificava a resposta dada pelo aparelho. Uma falha indica que o resultado dado não era igual ao esperado. Já um sucesso representa que o celular deu a resposta certa ao comando do áudio (ex: se o comando era "ligar a wi-fi"e ela foi realmente ligada).

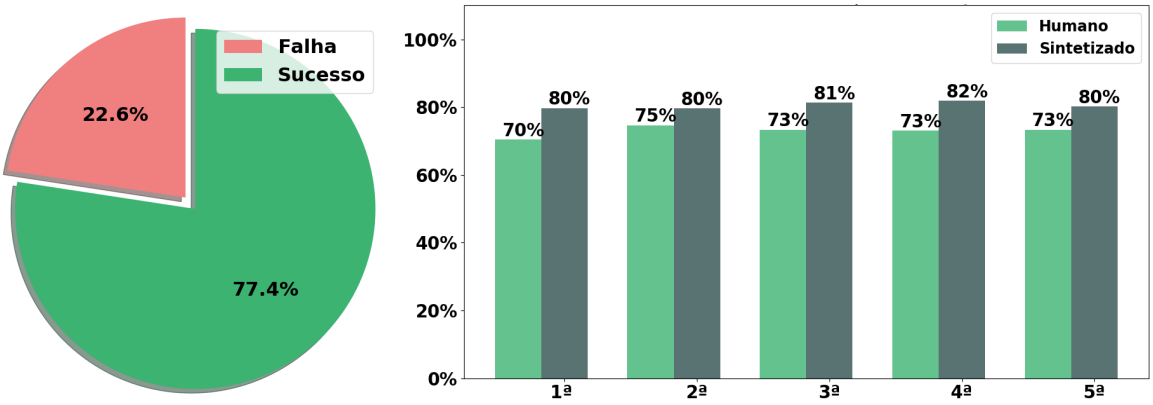
5.2.2.2 Resultados

Antes de mais nada, é importante ter em mente que sucesso e falha neste contexto (de transcrição) estão relacionados à transcrição correta e incorreta, respectivamente. A figura 43 tem dois gráficos. No primeiro, está porcentagem média de falha e sucesso das cinco execuções no teste de transcrição (43a). Tais resultados são referentes tantos aos áudios sintetizados, quanto humanos. 77,4% das transcrições foram reconhecidas corretamente e o restante não.

No segundo gráfico (43b) estão as porcentagens de sucesso dos áudios humanos e sintetizados por rodada, da primeira à quinta. As taxas referentes aos áudios sintetizados variam pouco, de 80% a 82%. Já os humanos tiveram taxa de sucesso de transcrição inferior aos sintetizados em todas as rodadas, variando de 70% a 75%.

Os resultados apresentados mostram que os áudios sintetizados têm uma porcentagem menor de falha, do que os humanos, relacionadas à transcrição (isso pode ser decorrente da ausência de ruído nos áudios sintéticos). No entanto, tal porcentagem não é tão inferior

assim. A maior diferença é de 10%, que é referente aos resultados da primeira rodada. Já a menor é de 5%, segunda rodada.



(a) Geral

(b) Por tipo e execução

Figura 43 – Taxa de falha e sucesso geral e por execução

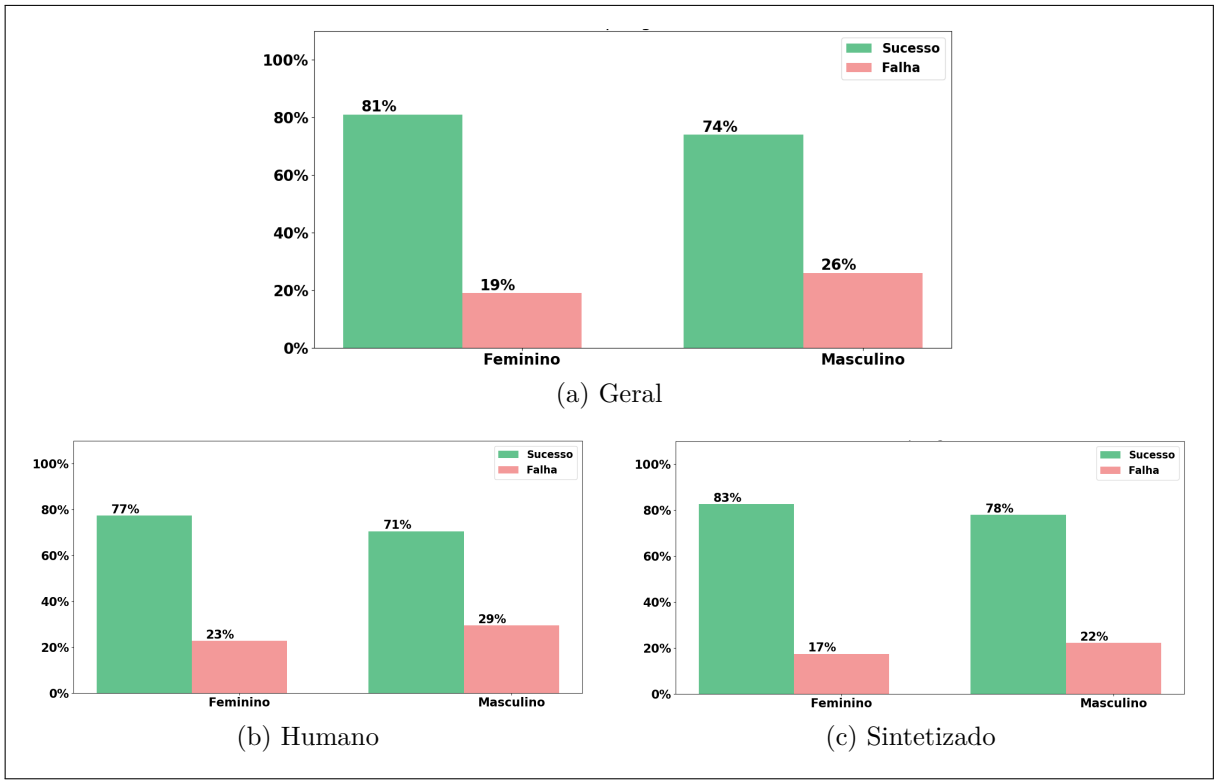
O objetivado nesta avaliação era que a taxa de sucesso de transcrição dos áudios sintetizados fosse realmente próxima das humanas. E elas são, tendo como média de taxa de sucesso dos áudios sintéticos e humanos de 80,6% e 73%, respectivamente. Isso mostra que dada uma variedade de áudios sintéticos com diferentes vozes, sotaques e gêneros, o comportamento do sistema em relação ao reconhecimento da fala sintética vai ser semelhante às variações da fala humana.

A Figura 44 mostra as taxas de sucesso em relação às cinco execuções dos testes de transcrição. É possível ver que, no geral (Figura 44a), a acurácia das transcrições dos áudios masculinos foi inferior aos femininos, com 81% e 74%, respectivamente. Analisando os gravados e sintéticos separadamente (figuras 44b e 44c, observa-se que o mesmo padrão foi seguido.

Vendo separadamente as taxas de sucesso por serviço na figura 45, observa-se que a que teve a maior acurácia foi o da Watson, com 87%, se distanciando em 14% da média dos humanos, que é de 73%, conforme visto.

Azure, Google e Polly tiveram taxas de acerto semelhantes, 79%, 80% e 81%, respectivamente. Desses, o que teve acurácia mais distante da média de transcrição corretas dos gravados foi o Polly, com diferença de 8%. Já a menor foi da Azure, com 6%.

Foram analisadas também as taxas de transcrição referentes aos gêneros de cada serviço (Figura 46). Em todos, as referentes aos áudios masculinos são menores que os femininos. O que chega mais próximo ao dos humanos masculinos é o da Google.



(a) Geral

(b) Humano

(c) Sintetizado

Figura 44 – Taxa de falha e sucesso por gênero geral (nas cinco execuções)

Os que tiveram maior taxa de acerto foram os da Watson, tanto os femininos, quanto os masculinos, com 88% e 86%, respectivamente. Esses valores se distanciam em 11% e 14% dos humanos de respectivos gêneros.

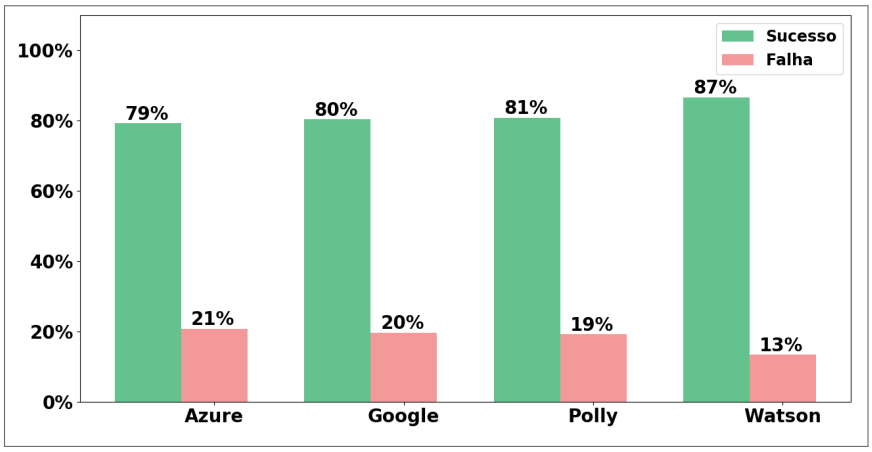
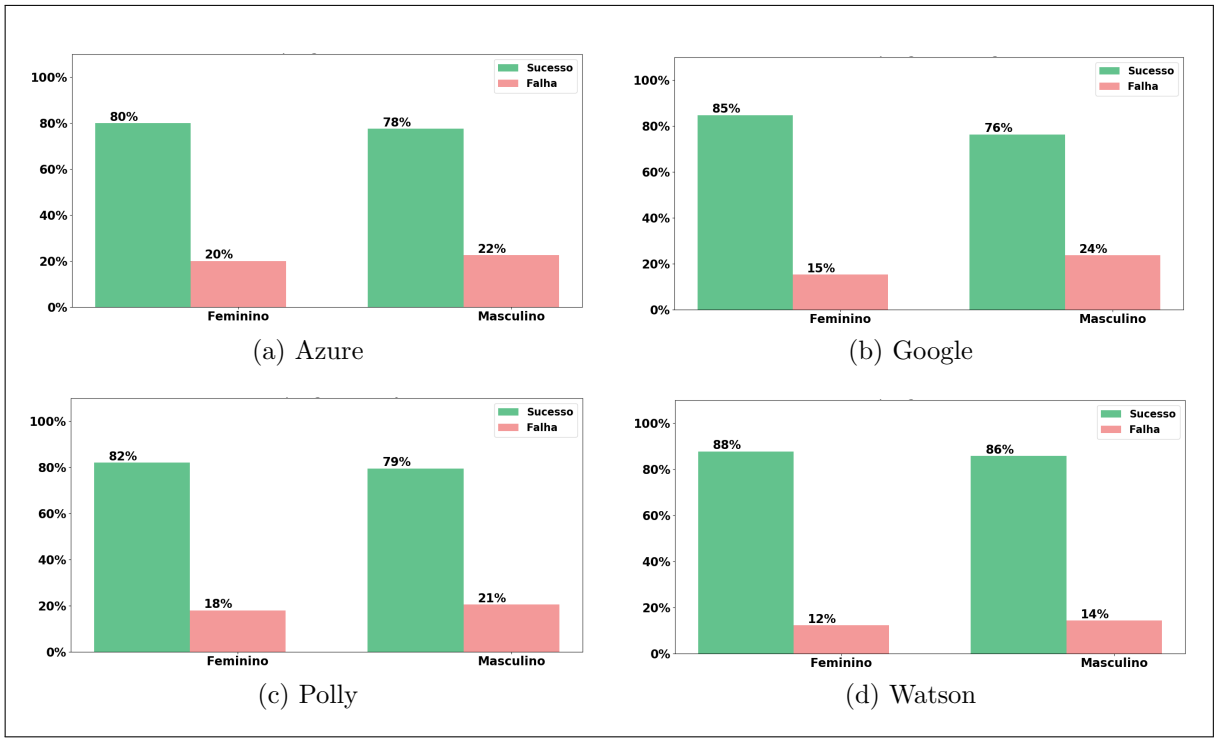


Figura 45 – Taxa de falha e sucesso por serviço

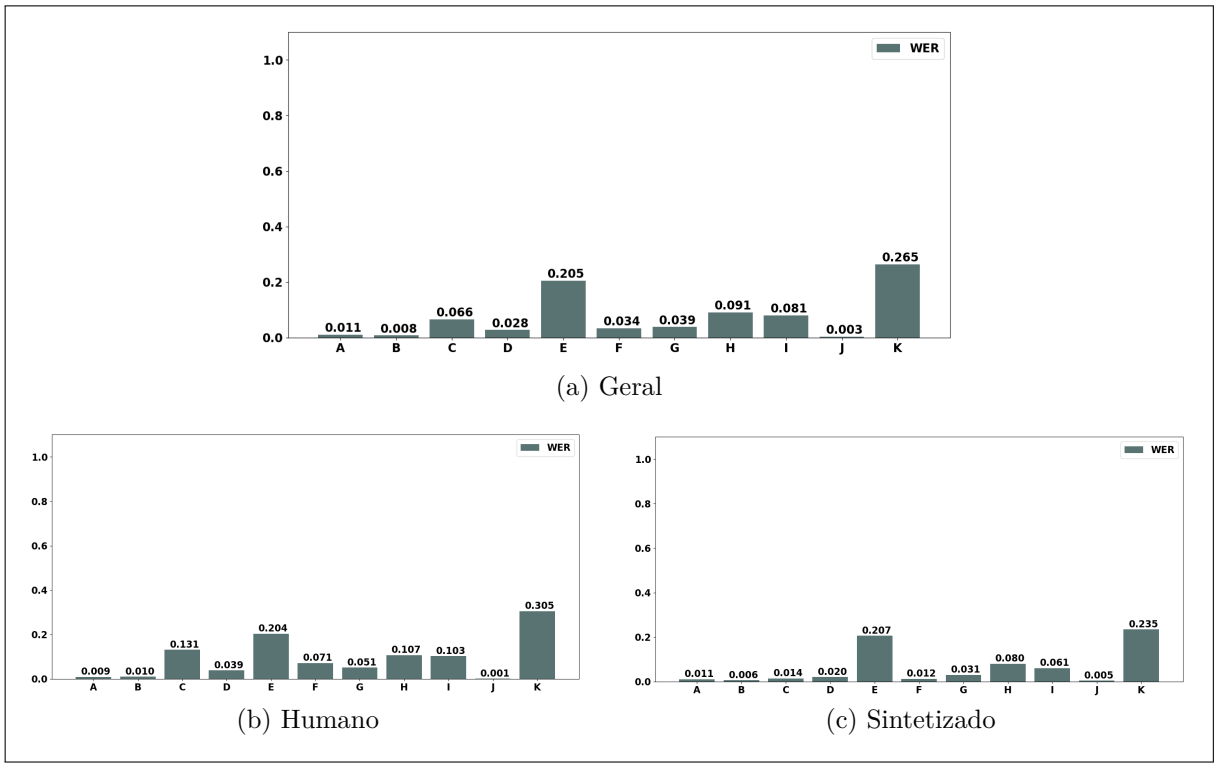
A taxa de erro de palavras (WER) também foi calculada. Os resultados são mostrados nos três gráficos da Figura 47. O primeiro (Figura 47a) apresenta a porcentagem geral da WER por sentença (A - K) referentes aos áudios humanos e sintéticos. As que tiveram maior erro foram a E e a K. Os outros gráficos mostram a WER dos áudios humanos e sintéticos separadamente, com maior porcentagem nas mesmas sentenças.



(a) Azure (b) Google

(c) Polly (d) Watson

Figura 46 – Taxa de falha e sucesso por serviço por gênero



(a) Geral

(b) Humano (c) Sintetizado

Figura 47 – WER por sentença

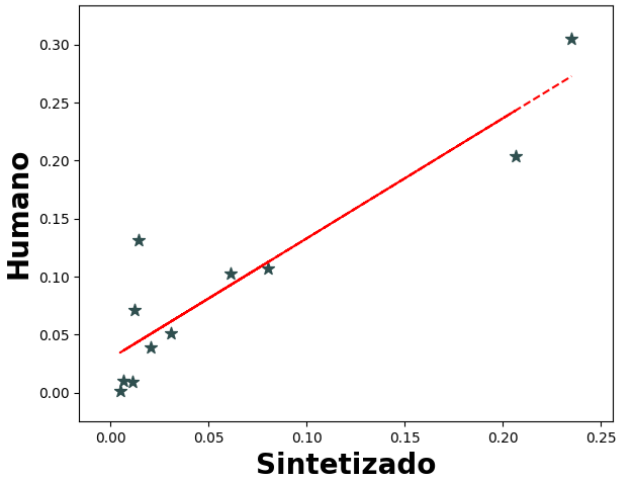
Tendo tais resultados (Figuras 47b e 47c), foi calculada a correlação de Pearson das taxas de erro das sentenças, referentes aos áudios humanos e sintetizados. Na figura 48

tem a tabela como valores da taxa de erro de cada sentença e o gráfico de dispersão com as mesmas.

A correlação desses dois conjuntos de dados foi calculada, tendo como resultado 0.914. Dado que o maior valor possível resultante dessa medida é 1 e valores acima de 0.9 são considerados de correlação muito forte, tem-se que o grau de correlação entre os dois conjuntos de dados avaliados é significativa. Assim, se porventura houver uma melhora no reconhecimento, pelo sistema, da frase K, usando áudios humanos, a probabilidade de que ele reconheça melhor os áudios sintéticos, é alta. A Figura 48 tem com mais detalhes os números referentes às taxas de erro das sentenças referentes aos áudios sintetizados e humanos, bem como o gráfico de dispersão das mesmas.

Sentença	Sintetizado	Humano
A	0,00948	0,01111
B	0,01000	0,00633
C	0,13114	0,01422
D	0,03886	0,02042
E	0,20358	0,20651
F	0,07147	0,01247
G	0,05107	0,03096
H	0,10684	0,08030
I	0,10254	0,06143
J	0,00121	0,00488
K	0,30525	0,23516

(a) WER por sentença



(b) Gráfico de dispersão das WERs por sentença

Figura 48 – Correlação das WERs dos áudios humanos e sintéticos

Além da correlação feita entre os dois conjuntos de dados de WERs, foi realizado um teste de hipótese chamado teste t de *Student*. Considerando a hipótese nula de que a média das WERs dos áudios sintetizados é a mesma que a das WERs dos humanos (gravados), o teste mostrou que há significância estatística entre os dois conjuntos de dados para o nível de significância de 0,05. Isso mostra uma forte evidência contra a hipótese nula, uma vez que essa hipótese foi rejeitada, onde o p-value correspondeu em cerca de 1,957%.

A Figura 49 apresenta a relação entre a qualidade dos áudios e o sucesso na execução dos testes automáticos de transcrição. A taxa de transcrições corretas dos áudios de menor qualidade (1) é a menor. Ela é ascendente até a 3 e depois mantém-se com valor de $\approx$ 64% nas qualidades 3, 4 e 5.

Conforme dito, além do teste de transcrição, foram realizados funcionais. Nestes últimos sucesso e falha dizem respeito à resposta dada pelo telefone depois da execução de determinado caso de teste (reprodução de um áudio). Uma falha pode ocorrer por vários

motivos, não necessariamente só por um defeito no celular. Algumas das causas das falhas são: erro de ambiente de teste, de setup, de script, de transcrição e da aplicação, estes últimos poderiam ser caracterizados como bug. Assim, as taxas de falha da Figura 50a são referentes a qualquer tipo de situação que impediu que o teste fosse bem sucedido. A situação é a mesma em todas as análises referentes a este tipo de teste (funcional). Aqui, não foi analisado caso a caso.

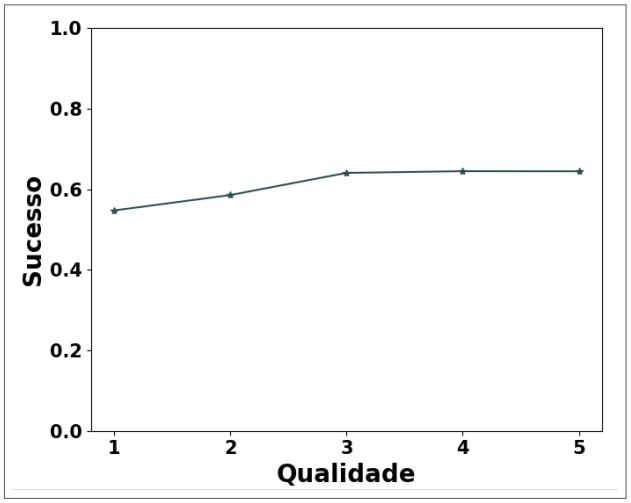
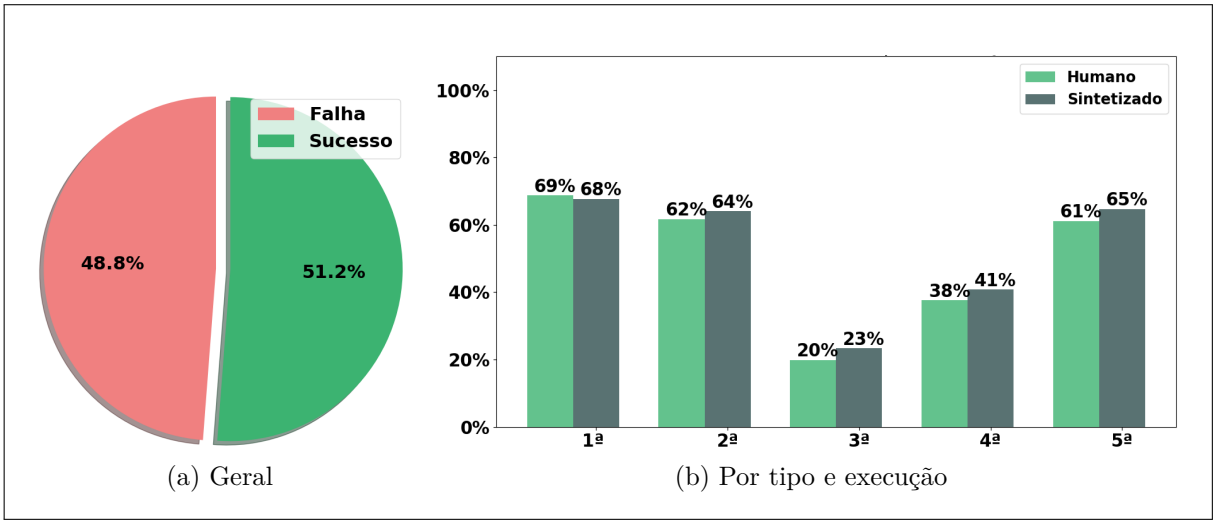


Figura 49 – Qualidade x Sucesso Automático (transcrição)

A Figura 50 mostra as taxas de sucesso (humano e sintetizado) por rodada, uma vez que, cada áudio foi executado cinco vezes, como no teste de transcrição. As porcentagens do sucesso sintetizado são próximas às dos humanos em todas as execuções, a maior diferença é de 4% (última rodada).

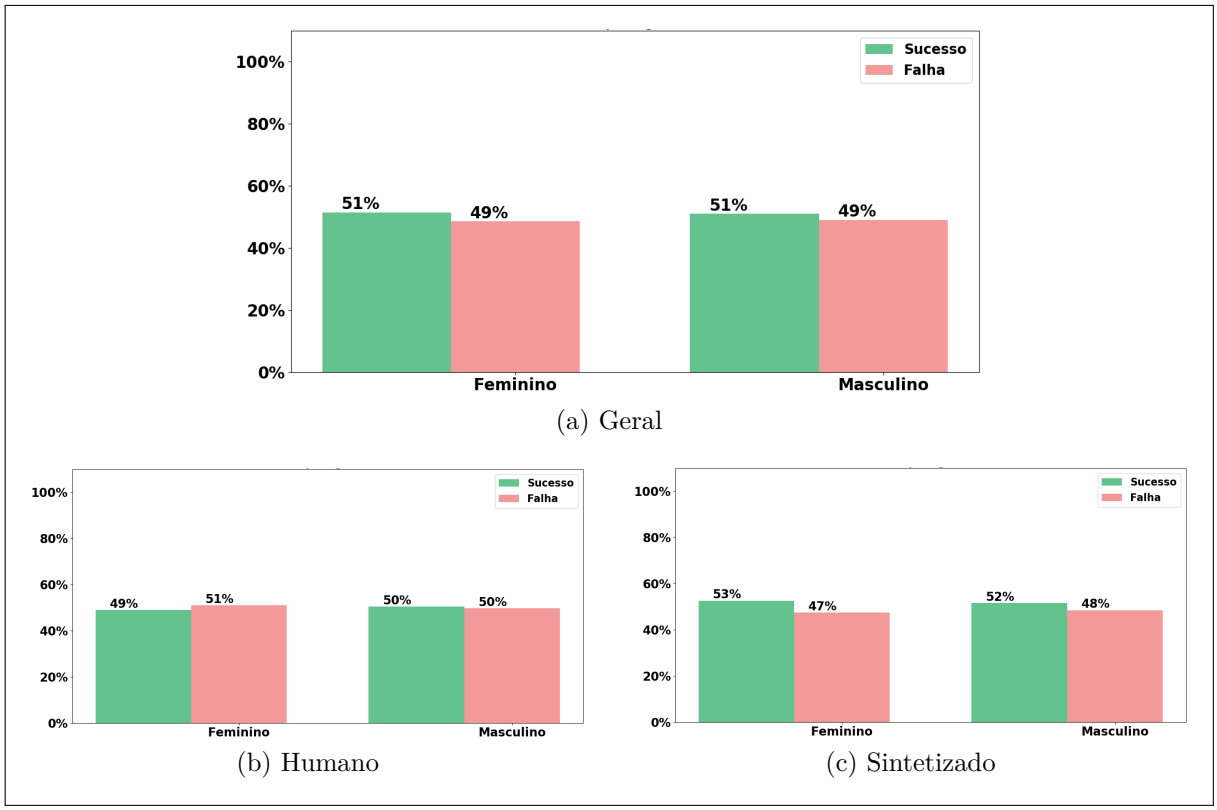


(a) Geral

(b) Por tipo e execução

Figura 50 – Taxa de falha e sucesso geral e por execução (funcional)

As taxas de sucesso e falha divididas por gênero são balanceadas, em torno de 50% (Figura 51). Em todos os casos, a taxa de falha é inferior ou igual à de sucesso, exceto do áudios femininos humanos, como mostra a Figura 51b.



(a) Geral

(b) Humano

(c) Sintetizado

Figura 51 – Taxa de falha e sucesso por gênero geral (nas cinco execuções)

A Figura 52 mostra um gráfico com as taxas de falha e sucesso do teste funcional dividido por serviço. Tais taxas são equilibradas em todos, em torno de 50% cada, exceto o da Watson, que teve maior uma diferença de 18% entre falha e sucesso.

Analisando separadamente as porcentagens de sucesso por gênero de cada serviço (Figura 52, observa-se que, novamente, elas são consideravelmente balanceadas nos da Azure, Google e Polly, sendo que a maior diferença é entre as taxas de falha e sucesso dos áudios femininos da Google e masculinos da Polly, com 8% de diferença em cada.

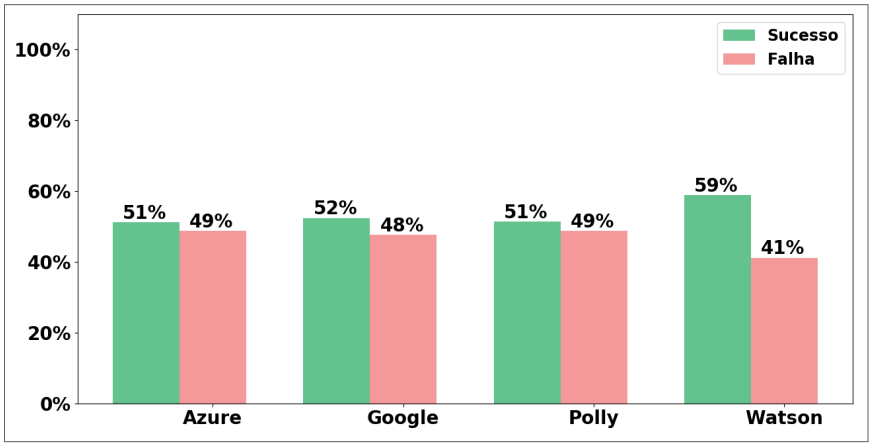


Figura 52 – Taxa de falha e sucesso por serviço

Já no serviço Watson as tais diferenças são maiores, tantos nos masculinos quantos

nos femininos, com 12% e 34% respectivamente, o que se distancia das mesmas taxas dos áudios humanos (Figura 51b).

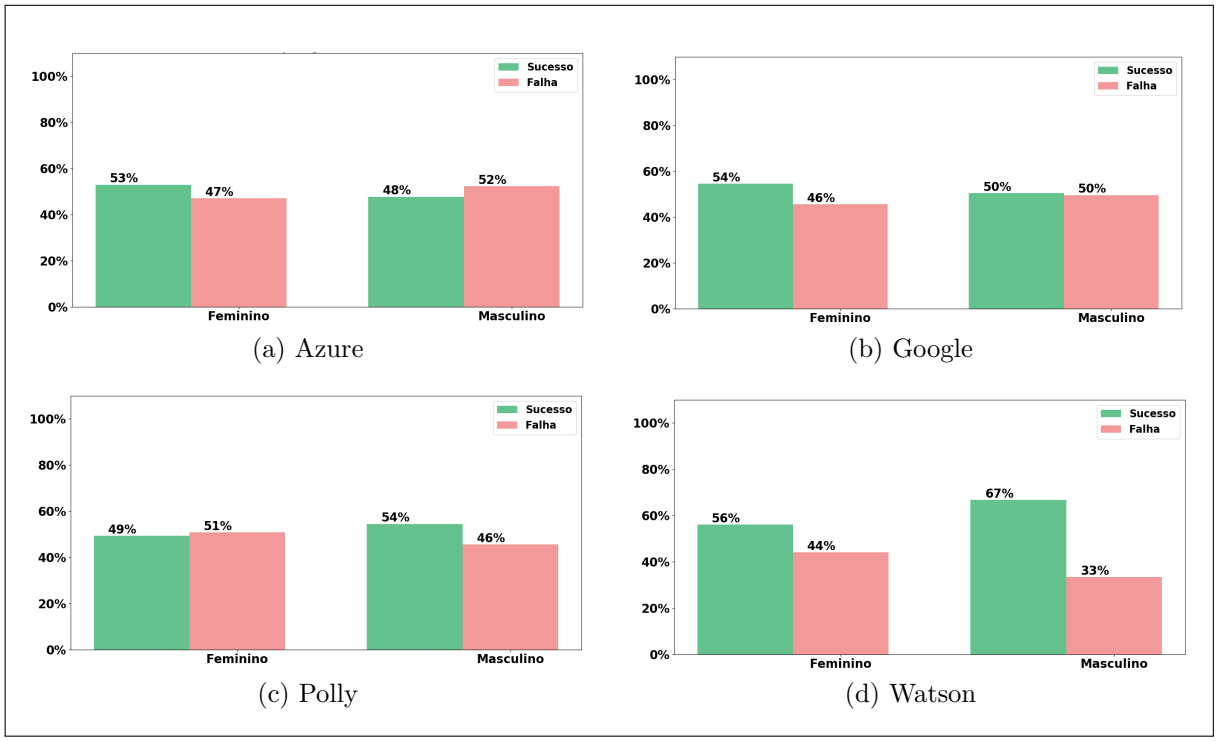


Figura 53 – Taxa de falha e sucesso por serviço por gênero

5.3 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os protocolos e resultados de dois experimentos. Um inicial, que apesar de ter sido relativamente pequeno com poucos dados, deu resultados que serviram de estímulo para o aprofundamento de pesquisa e desenvolvimento deste trabalho. E outro maior e mais completo, com mais dados avaliados.

Na avaliação objetiva, os resultados mostraram que os áudios masculinos sintetizados são mais confundidos com os humanos, ou seja, no geral, são mais naturais que os femininos. Dentre os quatro serviços de sintetização avaliados, o da Google foi o que teve melhor desempenho nos testes de turing e qualidade.

Sobre a avaliação subjetiva ainda do segundo experimento, foi possível ver que a porcentagem de transcrições corretas das falas sintetizadas, que feitas pelos sistema de reconhecimento do celular, foi parecido com das humanas. Além do mais, foi utilizada a métrica WER para calcular a taxa de erro de cada sentença. Depois disso, foi feita a correlação dos resultados das WERs das sentenças humanas e sintetizadas, que é muito alta.

6 CONCLUSÕES E TRABALHOS FUTUROS

Neste capítulo, discutimos, de forma geral, sobre as contribuições deste trabalho e sugestões para pesquisas futuras. O objetivo principal era desenvolver uma arquitetura de teste de sistema de reconhecimento de fala usando áudios sintetizadas. Tal arquitetura é composta de cinco módulos: processamento de sentenças, geração de sentenças equivalentes, filtragem de sentenças, sintetização da fala e teste de sistema de reconhecimento da fala.

6.1 CONTRIBUIÇÕES

Um dos principais desafios de tal proposta era validar se as falas sintéticas poderiam substituir as gravadas (humanas), e fazer um sistema para uso em ambiente real de teste. Para fazer tal validação, foram realizados experimentos e análises subjetivas (a partir da opinião de pessoas) e objetivas (com a execução de testes automáticos em um sistema STT) em relação a tais falas (sintetizadas).

O trabalho desenvolvido já vem sendo utilizado em ambiente real de teste (em *smartphones*) nas dependências do projeto CIn/Motorola. Ele tem contribuído tanto na economia de tempo, quanto financeira, principalmente em relação ao uso das falas sintetizadas como uma alternativa às gravadas (humanas).

Segundo um engenheiro (de teste) da equipe responsável por executar tais tipos de teste, para gravar cada sentença, uma pessoa leva, em média, 10 minutos (sem pós-processamento). Já para sintetizar, o tempo é em milissegundos. Então, considerando os áudios sintetizados para o experimento II, com o serviço da Google, o tempo de geração foi de 10,4 minutos referente a 312 áudios. Se esse processo fosse feito de forma gravada, seriam necessárias 52 horas para gravar a mesma quantidade de áudios.

6.2 TRABALHOS FUTUROS

Com os experimentos feitos, foi possível analisar uma série de fatores, mas ainda há bastante cenários que podem ser explorados, bem como melhoria de alguns módulos:

- **Experimentos com ruído nos áudios:** esta ideia surgiu porque acredita-se que um dos motivos que diferem os áudios humanos dos sintéticos é a ausência de ruído nesses últimos. Esse pensamento foi confirmado com um dos feedbacks dado por um dos voluntários que participou dos experimentos. Ele falou que ficou menos difícil de julgar alguns áudios sintéticos porque eles não têm ruído. Então, quando ele escutava e tinha dúvida em relação à fala, ele observava o nível de ruído e, a partir daí, julgava.

- **Experimentos com falas em outra linguagem**: como a linguagem explorada neste trabalho foi o inglês, outro aspecto a ser investigado é a sintetização em outras línguas (e.g: Português, Espanhol, Francês), que pode ser feita com os serviços aqui explorados. Os experimentos feitos para os áudios em inglês poderiam ser replicados para essas outras linguagens.

- **Método mais eficiente de geração de sentenças equivalentes:** Em relação à este tópico (que não foi o foco deste trabalho), pode ser feito um estudo mais aprofundado e aplicação de técnicas mais eficientes, utilizado aprendizagem de máquina. Uma vez que, este trabalho utilizou uma abordagem relativamente simples, baseada em sinônimos. Desta forma, não são geradas sentenças com estrutura sintática diferente da original e sabe-se que duas sentenças equivalentes não necessariamente têm a mesma estrutura sintática, mas sim, semântica. Com isso, possivelmente seja necessário utilizar uma nova técnica de fazer filtragem de sentenças e dependendo do método de geração e da aplicação, nem seja necessário filtrar.

REFERÊNCIAS

ACERO, A.; HON, H.-W.; HUANG, X. Hmm-based smoothing for concatenative speech synthesis. In: *Proc. of the Int. Conf. on Spoken Language Processing*. [S.l.: s.n.], 1998.

AWS, A. W. S. Amazon polly. In: . [s.n.], 2018. Acessado em: 22/12/2018. Disponível em: <https://aws.amazon.com/polly/>.

BIRKHOLZ, P.; MARTIN, L.; WILLMES, K.; KRöGER, B. J.; NEUSCHAEFER-RUBE, C. The contribution of phonation type to the perception of vocal emotions in german: An articulatory synthesis study. In: *The Journal of the Acoustical Society of America*. [S.l.: s.n.], 2015. p. 1503–1512.

CHANDU, K. R.; RALLABANDI, S. K.; SITARAM, S.; BLACK, A. W. Speech synthesis for mixed-language navigation instructions. In: . [S.l.]: ISCA, 2017.

COHEN, J.; BARTON, W.; PLOUMIS, J.; ELY, D. Automated testing of voice recognition software. In: . [s.n.], 2005. Disponível em: <https://patents.google.com/patent/US20050197836A1/>.

CREPY, H.; KUSNITZ, J. A.; LEWIS, B. Testing speech recognition systems using test data generated by text-to-speech conversion. In: . [s.n.], 2003. Disponível em: <https://patents.google.com/patent/US6622121B1/>.

ERRATTAHI, R.; HANNANI, A. E.; OUAHMANE, H. Automatic speech recognition errors detection and correction: A review. In: *International Conference on Natural Language and Speech Processing, ICNLSP*. [S.l.: s.n.], 2015. p. 32–37. ISSN 1877-0509.

FERNANDEZ, R.; RENDEL, A.; RAMABHADRAN, B.; HOORY, R. Using deep bidirectional recurrent neural networks for prosodic-target prediction in a unit-selection text-to-speech system. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. [S.l.: s.n.], 2016. p. 1099–1108.

GANDHI, S.; JAISWAL, P.; MOORE, V.; TOON, G. Determining speech recognition accuracy. In: . [s.n.], 2004. Disponível em: <https://patents.google.com/patent/US20040015350A1/>.

GONZALVO, X.; TAZARI, S.; CHAN, C. an; BECKER, M.; GUTKIN, A.; SILEN, H. Recent advances in google real-time hmm-driven unit selection synthesizer. In: . [S.l.]: ISCA, 2016.

GOOGLE. Cloud text-to-speech. In: . [s.n.], 2018. Acessado em: 22/12/2018. Disponível em: <https://cloud.google.com/text-to-speech/>.

IBM. Watson text to speech. In: . [s.n.], 2018. Acessado em: 22/12/2018. Disponível em: <https://text-to-speech-demo.ng.bluemix.net/>.

KHAN, R. A.; CHITODE, J. S. Concatenative speech synthesis: A review. In: *International Journal of Computer Applications (0975 – 8887)*. [S.l.: s.n.], 2016.

KLATT, D. H. Review of text-to-speech conversion for english. In: *The Journal of the Acoustical Society of America*. [S.l.: s.n.], 1987.

KRöGER, B. J. Minimal rules for articulatory speech synthesis. In: *Proceedings of EUSIPCO*. [S.l.: s.n.], 1992. p. 331–334.

KRöGER, B. J.; BIRKHOLZ, P. Articulatory synthesis of speech and singing: State of the art and suggestions for future research. In: *Department of Phoniatrics, Pedaudiology, and Communication Disorders, University Hospital Aachen and Aachen University*. [S.l.: s.n.], 2009. p. 306–319.

LEMMETTY, S. Review of speech synthesis technology. In: *Helsinki University of Technology. Department of Electrical and Communications Engineering*. [S.l.: s.n.], 1999.

LING, Z.; DENG, L.; YU, D. Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis. In: . [S.l.: s.n.], 2013. v. 21, p. 2129–2139.

LUKOSE, S.; UPADHYA, S. S. Text to speech synthesizer-formant synthesis. In: *International Conference on Nascent Technologies in the Engineering Field (ICNTE-2017)*. [S.l.: s.n.], 2017. ISBN 9781509027941.

MANASWI, N. K. In: *Deep Learning with Applications Using Python*. Apress, 2018. ISSN 2214-7853. Disponível em: <https://doi.org/10.1007/978-1-4842-3516-4>.

MICROSOFT. Azure text to speech. In: . [s.n.], 2018. Acessado em: 22/12/2018. Disponível em: <https://azure.microsoft.com/en-us/services/cognitive-services/text-to-speech/>.

OORD, A. V. D.; DIELEMAN, S.; ZEN, H.; SIMONYAN, K.; VINYALS, O.; GRAVES, A.; KALCHBRENNER, N.; SENIOR, A.; KAVUKCUOGLU, K. Wavenet: A generative model for raw audio. In: . [S.l.: s.n.], 2016.

OORD, A. van den; LI, Y.; BABUSCHKIN, I.; SIMONYAN, K.; VINYALS, O.; KAVUKCUOGLU, K. Parallel wavenet: Fast high-fidelity speech synthesis. In: . [S.l.: s.n.], 2017.

PALO, P. A review of articulatory speech synthesis. In: *Espoo*. [S.l.: s.n.], 2006.

PITRELLI, J. F.; BAKIS, R.; EIDE, E. M.; FERNANDEZ, R.; HAMZA, W.; PICHENY, M. A. The ibm expressive text-to-speech synthesis system for american english. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. [S.l.: s.n.], 2006. p. 1099–1108.

RAITIO, T.; SUNI, A.; YAMAGISHI, J.; PULAKKA, H.; NURMINEN, J.; VAINIO, M.; ALKU, P. Hmm-based speech synthesis utilizing glottal inverse filtering. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. [S.l.: s.n.], 2011. p. 153–165.

RESEARCH-HAIFA, I. Ibm virtual voice creator. In: . [s.n.], 2019. Acessado em: 19/04/2019. Disponível em: <http://www.research.ibm.com/haifa/dept/imt/ivvc.shtml#Contact>.

SHEN, J.; PANG, R.; WEISS, R. J.; SCHUSTER, M.; JAITLY, N.; YANG, Z.; CHEN, Z.; ZHANG, Y.; WANG, Y.; SKERRY-RYAN, R.; SAUROUS, R. A.; AGIOMYRGIANNAKIS, Y.; WU, Y. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In: . [S.l.: s.n.], 2018.

SIDDHI, D.; VERGHESE, J. M.; BHAVIK, D. Survey on various methods of text to speech synthesis. In: *International Journal of Computer Applications (0975 – 8887)*. [S.l.: s.n.], 2017.

SORIN, A.; SHECHTMAN, S.; RENDEL, A. Semi parametric concatenative tts with instant voice modification capabilities. In: *Proc. Interspeech 2017*. [s.n.], 2017. p. 1373–1377. Disponível em: <http://dx.doi.org/10.21437/Interspeech.2017-1202>.

TABET, Y.; BOUGHAZI, M. Speech synthesis techniques. a survey. In: *7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA)*. [S.l.: s.n.], 2011. p. 67–70.

THOMAS, S.; RAO, M. N.; MURTHY, H. A.; RAMALINGAM, C. Natural sounding tts based on syllable-like units. In: *14th European Signal Processing Conference*. [S.l.: s.n.], 2006.

TIOMKIN, S.; MALAH, D.; SHECHTMAN, S.; KONS, Z. A hybrid text-to-speech system that combines concatenative and statistical synthesis units. In: *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING*. [S.l.: s.n.], 2011. p. 1278–1288.

VIJAYARANI, D. S.; MR.S.DHAYANAND. Liver disease prediction using svm and naïve bayes algorithms. In: *International Journal of Science, Engineering and Technology Research (IJSETR)*. [S.l.: s.n.], 2015.

VIMALA, C.; RADHA. A review on speech recognition challenges and approaches. In: *World of Computer Science and Information Technology Journal (WCSIT)*. [S.l.: s.n.], 2012. p. 1–7. ISSN 2221-0741.

W3C. Speech synthesis markup language (ssml) version 1.1. In: . [s.n.], 2018. Acessado em: 22/12/2018. Disponível em: <https://www.w3.org/TR/speech-synthesis11/>.

YANG, J.; HONG, Z. Subjective evaluation of speech recognition in noise. In: *Journal of Applied Science and Engineering Innovation Vol.2 No.2*. [S.l.: s.n.], 2015.

YU, D.; DENG, L. In: *Automatic Speech Recognition*. [S.l.]: Springer, 2015. ISSN 1860-4862.

ZEN, H.; AGIOMYRGIANNAKIS, Y.; EGBERTS, N.; HENDERSON, F.; SZCZEPANIAK, P. Fast, compact, and high quality lstm-rnn based statistical parametric speech synthesizers for mobile devices. In: . [S.l.: s.n.], 2016.

ZEN, H.; SENIOR, A.; SCHUSTER, M. Statistical parametric speech synthesis using deep neural networks. In: *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*. [S.l.: s.n.], 2013. p. 7962–7966.

ZEN, H.; TOKUDA, K.; BLACK, A. W. Statistical parametric speech synthesis. In: *Speech Communication 51*. [S.l.: s.n.], 2009. p. 1039–1064.

APÊNDICE A – VOZES DISPONÍVEIS NOS SERVIÇOS DE SÍNTESE

A.1 AMAZON POLLY

Tabela 13 – Vozes disponíveis no Amazon Polly

Idioma	**Nome(s) Feminino(s)**	**Nome(s) Masculino(s)**
Alemão	Marlene Vicki	Hans
Coreano	Seoyeon	
Dinamarquês	Naja	Mads
Espanhol Europeu	Conchita	Enrique
Espanhol Latino-Americano	Penélope	Miguel
Francês	Céline/Celine	Mathieu
Francês Canadense	Chantal	
Galês	Gwyneth	
Hindi	Aditi (bilíngue com Inglês Indiano)	
Holandês	Lotte	Ruben
Inglês Australiano	Nicole	Russell
Inglês Britânico	Amy Emma	Brian
Inglês Indiano	Aditi (bilíngue com hindi) Raveena	
Inglês Americano	Ivy Joanna Kendra Kimberly Salli	Joey Justin Matthew
Inglês Galês		Geraint
Islandês	Dora	Karl
Italiano	Carla	Giorgio
Japonês	Mizuki	Takumi
Norueguês	Liv	
Polonês	Ewa Maja	Jacek Jan
Português Brasileiro	Vitória	Ricardo
Português Europeu	Inês	Cristiano
Romeno	Carmen	
Russo	Tatyana	Maxim
Sueco	Astrid	
Turco	Filiz	

Fonte: (AWS, 2018)

A.2 GOOGLE *TEXT TO SPEECH*

Tabela 14 – Vozes disponíveis no Google *Text-to-Speech*

Idioma	**Tipo de Voz**	**Voz(es) Feminina(s)**	**Voz(es) Masculina(s)**
Alemão	Padrão	de-DE-Standard-A	de-DE-Standard-B
	WaveNet	de-DE-Wavenet-A	de-DE-Wavenet-B
		de-DE-Wavenet-C	de-DE-Wavenet-D
Coreano	Padrão	ko-KR-Standard-A	
	WaveNet	ko-KR-Wavenet-A	
Espanhol	Padrão	es-ES-Standard-A	
Francês	Padrão	fr-FR-Standard-A	fr-FR-Standard-B
		fr-FR-Standard-C	fr-FR-Standard-D
	WaveNet	fr-FR-Wavenet-A	fr-FR-Wavenet-B
		fr-FR-Wavenet-C	fr-FR-Wavenet-D
Francês Canadense	Padrão	fr-CA-Standard-A	fr-CA-Standard-B
		fr-CA-Standard-C	fr-CA-Standard-D
Holandês	Padrão	nl-NL-Standard-A	
	WaveNet	nl-NL-Wavenet-A	
Inglês Australiano	Padrão	en-AU-Standard-A	en-AU-Standard-B
		en-AU-Standard-C	en-AU-Standard-D
	WaveNet	en-AU-Wavenet-A	en-AU-Wavenet-B
		en-AU-Wavenet-C	en-AU-Wavenet-D
Inglês Britânico	Padrão	en-GB-Standard-A	en-GB-Standard-B
		en-GB-Standard-C	en-GB-Standard-D
	WaveNet	en-GB-Wavenet-A	en-GB-Wavenet-B
		en-GB-Wavenet-C	en-GB-Wavenet-D
Inglês Americano	Padrão	en-US-Standard-C	en-US-Standard-B
		en-US-Standard-E	en-US-Standard-D
	WaveNet	en-US-Wavenet-C	en-US-Wavenet-A
		en-US-Wavenet-E	en-US-Wavenet-B
		en-US-Wavenet-F	en-US-Wavenet-D
Italiano	Padrão	it-IT-Standard-A	
	WaveNet	it-IT-Wavenet-A	
Japonês	Padrão	ja-JP-Standard-A	
	WaveNet	ja-JP-Wavenet-A	
Português Brasileiro	Padrão	pt-BR-Standard-A	
Sueco	Padrão	sv-SE-Standard-A	
Turco	Padrão	tr-TR-Standard-A	

Fonte: (GOOGLE, 2018)

A.3 IBM WATSON *TEXT TO SPEECH*

Tabela 15 – Vozes disponíveis na IBM Watson

Idioma	**Nome(s) Feminino(s)**	**Nome(s) Masculino(s)**
Alemão	Birgit	Dieter
Espanhol Casteliano	Laura	Enrique
Espanhol Latino-Americano	Sofia	
Espanhol Norte-Americano	Sofia	
Francês	Renee	
Inglês Britânico	Kate	
Inglês Americano	Alisson Lisa	Michael
Italiano	Francesca	
Japonêes	Emi	
Português Brasileiro	Isabela	

Fonte: (IBM, 2018)

A.4 MICROSOFT AZURE *TEXT TO SPEECH*

Tabela 16 – Vozes disponíveis no Microsoft Azure Bing Speech

Idioma	**Nome(s) Feminino(s)**	**Nome(s) Masculino(s)**
Alemão (Austria)		(de-AT, Michael)
Alemão (Suiça)		(de-CH, Karsten)
Alemão (Alemanha)	(de-DE, Hedda) (de-DE, HeddaRUS)	(de-DE, Stefan, Apollo)
Árabe (Egito)	(ar-EG, Hoda)	
Árabe (Arábia Saudita)		(ar-SA, Naayf)
Búlgaro		(bg-BG, Ivan)
Catalão (Espanha)	(ca-ES, HerenaRUS)	
Checo		(cs-CZ, Jakub)
Chinês (China)	(zh-CN, HuihuiRUS) (zh-CN, Yaoyao, Apollo)	(zh-CN, Kangkang, Apollo)
Chinês (Hong Kong)	(zh-HK, Tracy, Apollo) (zh-HK, TracyRUS)	(zh-HK, Danny, Apollo)
Chinês (Taiwan)	(zh-TW, Yating, Apollo) (zh-TW, HanHanRUS)	(zh-TW, Zhiwei, Apollo)
Coreano	(ko-KR, HeamiRUS)	
Croata		(hr-HR, Matej)
Dinamarquês	(da-DK, HelleRUS)	
Espanhol (Espanha)	(es-ES, Laura, Apollo) (es-ES, HelenaRUS)	(es-ES, Pablo, Apollo)
Eslovaco		(sk-SK, Filip)
Esloveno		(sl-SI, Lado)
Espanhol (México)	(es-MX, HildaRUS)	(es-MX, Raul, Apollo)
Finlandês	(fi-FI, HeidiRUS)	
Francês (Canadá)	(fr-CA, Caroline) (fr-CA, Caroline)	
Francês (Suiça)		(fr-CH, Guillaume)
Francês (França)	(fr-FR, Julie, Apollo) (fr-FR, HortenseRUS)	(fr-FR, Paul, Apollo)
Grego		(el-GR, Stefanos)
Hebraico (Israel)		(he-IL, Asaf)
Hindi (Índia)	(hi-IN, Kalpana, Apollo) (hi-IN, Kalpana)	(hi-IN, Hemant)
Holandês	(nl-NL, HannaRUS)	
Hungaro		(hu-HU, Szabolcs)
Inglês Australiano	(en-AU, Catherine) (en-AU, HayleyRUS)	

Idioma	Nome(s) Feminino(s)	Nome(s) Masculino(s)
Inglês Canadense	(en-CA, Linda) ((en-CA, HeatherRUS)	
Inglês Britânico	(en-GB, Susan, Apollo) (en-GB, HazelRUS, Apollo)	(en-GB, George, Apollo)
Inglês (Irlanda)		(en-IE, Sean)
Inglês (Índia)	(en-IN, Heera, Apollo) (en-IN, PriyaRUS)	(en-IN, Ravi, Apollo)
Inglês (EUA)	(en-US, ZiraRUS) (en-US, JessaRUS) (en-US, Jessa24kRUS)	(en-US, BenjaminRUS) (en-US, Guy24kRUS)
Indonésio		(id-ID, Andika)
Italiano	(it-IT, Cosimo, Apollo)	(it-IT, LuciaRUS)
Japonês	(ja-JP, Ayumi, Apollo) (ja-JP, HarukaRUS)	(ja-JP, Ichiro, Apollo)
Malaio		(ms-MY, Rizwan)
Norueguês	(ms-MY, Rizwan)	
Polonês	(pl-PL, PaulinaRUS)	
Português (Brasil)	(pt-BR, HeloisaRUS)	(pt-BR, Daniel, Apollo)
Português (Portugal)	(pt-PT, HeliaRUS)	
Romeno		(ro-RO, Andrei)
Russo	(ru-RU, Irina, Apollo) (ru-RU, Pavel, Apollo)	(ru-RU, Pavel, Apollo)
Sueco	(sv-SE, HedvigRUS)	
Tâmil (Índia)		(ta-IN, Valluvar)
Télugo (Índia)	(te-IN, Chitra)	
Thailandês		(th-TH, Pattara)
Turco	(tr-TR, SedaRUS)	
Vietnamita		(vi-VN, An)

Fonte: (MICROSOFT, 2018)