UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA
Autorizada pelo Decreto Federal nº 77.496 de 27/04/76
Recredenciamento pelo Decreto nº17.228 de 25/11/2016
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
COORDENAÇÃO DE INICIAÇÃO CIENTÍFICA

XXVIII SEMINÁRIO DE INICIAÇÃO CIENTÍFICA DA UEFS
SEMANA NACIONAL DE CIÊNCIA E TECNOLOGIA - 2024

APERFEIÇOAMENTO DO SISTEMA DE RECONHECIMENTO DE COMANDOS DE VOZ

Igor Figueredo Soares[1]; Anfranserai Morais Dias[2]
1. Bolsista – PROBIC/UEFS, Graduando em Engenharia de Computação, UEFS, e-mail: ifs54@hotmail.com
2. Orientador, Departamento de Tecnologia, UEFS, e-mail: anfranserai@ecomp.uefs.br

PALAVRAS-CHAVE: comandos de voz; interface; DeepSpeech.

INTRODUÇÃO

Uma interface via voz, na linguagem do usuário, é a mais natural, flexível, eficiente, e econômica forma de comunicação (Ynoguti, 1999). O reconhecimento de fala por máquina é um desafio devido à natureza dinâmica e variável do sinal de fala. O presente trabalho visa o aperfeiçoamento de um sistema de reconhecimento de comandos por voz, inicialmente desenvolvido no projeto de Iniciação em Desenvolvimento Tecnológico e Inovação (PIBITI/CNPq) 2021/2022 (Júnior; Dias, 2022). O sistema aprimorado é composto por 2 módulos principais: o de transcrição e o de classificação. O projeto será usado no Robô Guia de Visitação ao LaboTec 3 (Guibô).

MATERIAL E MÉTODOS OU METODOLOGIA (ou equivalente)

O primeiro objetivo do trabalho foi a construção de um conjunto de dados (*dataset*) para o treinamento e avaliação de desempenho do sistema. Para isso, foi solicitado a voluntários a gravação de áudios pronunciando 120 frases duas vezes, em arquivos de áudios distintos. As frases representam comandos, nomes de espaços do prédio, além de perguntas gerais e específicas sobre os laboratórios. As frases foram definidas com base no ambiente que o robô atuará e nas ações que ele será capaz de executar. A Tabela 1 ilustra alguns exemplos dessas frases.

Tabela 1 - Frases do conjunto de dados criado para o modelo acústico

Frases			
onde estou	quem é você	como chego à sala	desligar ar condicionado
ligar luzes	preciso de ajuda	apresente a sala	deixe o caminho livre

O módulo de transcrição consiste em um reconhecedor de fala. Ele segue a mesma estrutura definida no reconhecedor de fala do trabalho anterior, composto por um alfabeto, um modelo acústico e um modelo de linguagem. O módulo de classificação recebe o texto transcrito e compara com a lista de frases pré-estabelecidas. Ele então identifica qual das frases é a mais semelhante ao texto transcrito, verificando a similaridade textual entre as sentenças. O módulo de classificação foi implementado utilizando o modelo BERTimbau Base (Souza; Nogueira; Lotufo, 2020).

Para a avaliação de desempenho dos modelos implementados, foram utilizadas as métricas *Word Error Rate* (WER), *Character Error Rate* (CER) e a quantidade de

frases que foram transcritas ou classificadas incorretamente (no caso do modelo de classificação).

RESULTADOS E/OU DISCUSSÃO (ou Análise e discussão dos resultados)

A coleta dos dados para a construção do *dataset* resultou em um total de 5367 gravações, sendo 12 vozes femininas e 14 vozes masculinas. Os voluntários gravaram cada uma das frases duas vezes em arquivos de áudio distintos. Os áudios foram gravados no formato WAV, a uma taxa de amostragem de 16 kHz, em canal único.

A documentação do DeepSpeech recomenda evitar o uso de áudios curtos. Para garantir o melhor desempenho do modelo, foi necessário padronizar os dados coletados duplicando e concatenando os arquivos. O conjunto de dados construído foi dividido em 3 subconjuntos, onde 1534 áudios para treinamento e 767 para validação. Para o *dataset* de teste, as gravações foram mantidas originais, totalizando 1328 áudios.

Modelos Acústicos Preliminares

Foram treinados quatro modelos distintos para avaliar qual deles produziria os melhores resultados. Os Modelo 1 e 2 foram treinados utilizando todos os áudios disponíveis no dataset do *Common Voice* (Ardila et al., 2019) na versão 16.1 (Mozilla Common Voice, 2024). Para o Modelo 3, foi realizada uma filtragem do conjunto de dados mantendo apenas os áudios que continham pelo menos uma das palavras presentes nas 120 frases escolhidas, desconsiderando artigos e preposições. No Modelo 4, permaneceram apenas os áudios que continham pelo menos duas das palavras presentes nas frases definidas. A Tabela 2 apresenta os parâmetros de treinamento dos modelos.

Tabela 2 - Parâmetros nos treinamentos dos modelos DeepSpeech preliminares.

Identificação	***epochs*[1]**	***learning rate***	***plateau reduction***	***es_epochs***	***plateau epochs***	***n_hidden***	***batch size***
Modelo 1	39	0.0001	0.001	12	4	2048	64
Modelo 2	100	0.001	0.1	-	4	2048	64
Modelo 3	39	0.0001	0.001	12	4	2048	64
Modelo 4	62	0.0001	0.001	12	4	2048	64

[1] Inicialmente, todos os modelos foram configurados para 100 epochs. O número efetivo de epochs foi reduzido devido ao uso do parâmetro early stopping.

O melhor desempenho foi obtido pelo Modelo 1 (ver Tabela 3), que apresentou um WER médio de 15,43%, CER médio de 10,76% e a função de perda (Loss) de 27,02. O segundo foi o do Modelo 3, que, apesar de apresentar um Loss superior aos demais, atingiu um WER e CER médio abaixo dos Modelos 2 e 4.

Tabela 3 - Desempenho dos modelos acústicos preliminares.

Identificação	**WER**	**CER**	**Loss**
Modelo 1	15,43%	10,76%	27,02
Modelo 2	29,20%	22,54%	41,59
Modelo 3	22,66%	16,34%	46,52
Modelo 4	29,58%	22,19%	45,56

Otimização dos Modelos

O *fine-tuning* permite ajustar um modelo pré-treinado a um conjunto de dados específicos para melhorar seu desempenho em uma tarefa específica. Dessa forma, os

modelos foram ajustados ao *dataset* coletado, mantendo os mesmos parâmetros utilizados no treinamento inicial (vide Tabela 2). O processo de otimização foi aplicado aos Modelos 1 e 3. Os testes foram conduzidos com o mesmo subconjunto de teste. Na Tabela 4 apresenta os resultados de desempenho dos modelos após a otimização.

Tabela 4 - Desempenho dos modelos acústicos após a otimização.

Identificação	WER	CER	Loss	epochs
Modelo 1 Otimizado	3,93%	2,23%	0,35	29
Modelo 3 Otimizado	4,08%	2,31%	1,21	43

Os resultados mostram uma redução no WER e no CER para ambos os modelos após o ajuste fino, o que indica uma melhoria no reconhecimento de palavras e caracteres. Embora os modelos tenham apresentado resultados bastante semelhantes, o Modelo 1 Otimizado demonstrou taxas de erro ligeiramente inferiores ao Modelo 3 Otimizado. O Modelo 1 foi selecionado como o modelo acústico para implementação no sistema.

Módulo de Classificação

Do conjunto de 1328 gravações de teste, 145 apresentaram algum erro na transcrição, resultando em comandos diferentes dos reais. O módulo de classificação visa diminuir esse erro. Dentro desse contexto, para validar a utilização do módulo as transcrições do *dataset* de teste foi submetida à classificação. A quantidade de gravações transcritas incorretamente foi reduzida de 145 para 54, que representa uma redução de 62,76%. A taxa de erro é de cerca de 4,07% no conjunto de testes.

Além disso, é importante destacar que nas frases que se referem à identificação de salas, como “sala s um” e suas variações, o desempenho do classificador pode ser limitado. Neste caso, a sentença é transcrita com a omissão do número que identifica a sala. Logo, o classificador não tem informações suficientes para inferir corretamente qual seria a sala, podendo atribuir qualquer uma das frases. Para contornar essa limitação, o sistema pode solicitar ao usuário que repita o número da sala.

Palavra de Ativação

No trabalho inicial, foi estabelecido a utilização do MyCroft para a execução dos comandos e abstração dos processos. O MyCroft é um assistente virtual de código aberto que permite uma série de customizações. Ele fornece as ferramentas para o treinamento e implementação da palavra de ativação personalizada. O termo escolhido como palavra de ativação foi “Guibô”. O conjunto de dados das gravações de vozes para a palavra de ativação foi construído de modo análogo ao *dataset* coletado. Além disso, foram incluídas algumas palavras começando com "gui" e outras terminando com "bo" ou "ô". As gravações dessas palavras foram feitas por 2 voluntários (uma voz masculina e uma voz feminina). Ao todo, foram obtidas 62 gravações com a palavra de ativação, 42 foram destinadas ao treinamento e 20 para testes. O grupo sem a palavra de ativação totalizou 281 áudios, dos quais 175 para treinamento e 106 para testes.

O treinamento da *wake word* foi executado utilizando 200 *epochs*, resultando em um *Loss* de treinamento de 0,0281 e um *Loss* de validação de 0,0691. Nos testes do modelo, do total das 20 gravações com a palavra de ativação, 15 foram reconhecidas corretamente, uma taxa de acerto de 75%. Já no conjunto de testes sem a palavra de ativação, dois áudios foram reconhecidos incorretamente, do total de 106, resultando em

uma taxa de erro de cerca de 1,89%.

CONSIDERAÇÕES FINAIS (ou Conclusão)

O sistema aprimorado para o reconhecimento de comandos é capaz de desempenhar as atribuições para as quais foi projetado. A incorporação do módulo classificador proporcionou ao sistema ganhos significativos na identificação das frases. Além disso, a personalização da palavra de ativação torna a interação do usuário com o robô guia mais intuitiva.

Contudo, o sistema ainda apresenta limitações na transcrição de algumas frases e na identificação da palavra de ativação em determinadas situações. Assim, em trabalhos futuros, o sistema pode ser aperfeiçoado, com o refinamento do modelo acústico desenvolvido e da palavra de ativação, utilizando conjuntos de dados maiores.

REFERÊNCIAS

ARDILA, R. et al. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*. Disponível em: https://doi.org/10.48550/arXiv.1912.06670. Acesso em: 31 ago. 2024.

MOZILLA. 2020. Mozilla DeepSpeech 0.9.3 documentation: *deepspeech model*. DeepSpeech Model. Disponível em: https://deepspeech.readthedocs.io/en/r0.9/DeepSpeech.html. Acesso em: 31 ago. 2024.

MOZILLA COMMON VOICE. 2024. *Datasets*. Disponível em: https://commonvoice.mozilla.org/pt/datasets. Acesso em: 31 ago. 2024.

OLIVEIRA JÚNIOR, W. L. F. D.; DIAS, A. M. 2022. Desenvolvimento de sistema para reconhecimento de comandos de voz para interface homem-máquina de um robô para guiar visitantes. *In*: Anais dos Seminários de Iniciação Científica, n. 26., Feira de Santana. Anais [...]. Feira de Santana: Universidade Estadual de Feira de Santana. Disponível em: https://doi.org/10.13102/semic.vi26.10514. Acesso em: 31 ago. 2024.

PDSOUND.ORG. *Public Domain Sounds Backup*. 2009. Gerenciado por Iwan Gabovitch. Disponível em: https://pdsounds.tuxfamily.org/. Acesso em: 31 ago. 2024.

REIMERS, N.; GUREVYCH, I. 2019. Sentence-BERT: sentence embeddings using siamese bert-networks. [S.L.]. *ArXiv*. Disponível em: http://dx.doi.org/10.48550/ARXIV.1908.10084. Acesso em: 31 ago. 2024.

SOUZA, F.; NOGUEIRA, R.; LOTUFO, R. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. *In*: Intelligent Systems: 9th Brazilian Conference, BRACIS 2020, Rio Grande, Brazil, October 20–23. Proceedings [...] , Part I 9. Springer International Publishing, p. 403-417. Disponível em: https://doi.org/10.1007/978-3-030- 61377-8_28. Acesso em: 31 ago. 2024.

YNOGUTI, C. A. 1999. Reconhecimento de fala contínua usando modelos ocultos de Markov. 138 p. Tese (doutorado) - Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação, Campinas, SP. Disponível em: https://doi.org/10.47749/T/UNICAMP.1999.175850. Acesso em: 31 ago. 2024.