Análise de Interrupções de Áudio em Processos de Reconhecimento de Fala Para Aplicações em Sistemas Elétricos de Potência

Victor H. Yoshizumi*. Sofia M. A. Lopes*. Danilo H. Spatti.**
Rogério A. Flauzino*. Ivan N. da Silva*. Ivan G. Ricci*. Alexandre G. C. Latorre*****

**Departamento de Engenharia Elétrica e de Computação,*
Escola de Engenharia de São Carlos, Universidade de São Paulo (EESC-USP),
São Carlos, São Paulo, Brasil (e-mail: yoshizumi@usp.br, sofia.moreira.lopes@usp.br,
raflauzino@usp.br, insilva@sc.usp.br).
*** Departamento de Sistemas de Computação,*
Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo (ICMC-USP),
São Carlos, São Paulo, Brasil (e-mail: spatti@icmc.usp.br).
****ARGO Transmissão de Energia S/A*
São Paulo, São Paulo, Brasil (e-mail: ivan.ricci@argoenergia.com.br, alexandre.latorre@argoenergia.com.br).

Abstract: Speech recognition as part of automatic decision-making systems has advanced in recent years, becoming a consistent reality in several engineering sectors. Especially in Power Systems, Speech-to-text recognition allows a significant increase in the quality of process operation involving audio communication. In this way, it becomes possible to transcribe the audios involving the communication and also future audits. A growing number of tools have been proposed lately in order to automate speech recognition, but these still have analysis limitations, not allowing redundancy in the transcription process, for example. This work proposes a methodology for analyzing audio in separate channels based on recordings of calls between electrical system operators, aiming increasing the degree of robustness in the application of speech-to-text recognition processes.

Resumo: O reconhecimento de fala como parte de mecanismos automáticos de auxílio à tomada de decisão tem avançado bastante nos últimos anos, tornando-se uma realidade consistente em vários setores da engenharia. Especialmente em Sistemas Elétricos de Potência, o reconhecimento do tipo Fala-para-texto permite um incremento significativo da qualidade na operação em processos que envolvem comunicação via áudio. Desta forma, torna-se possível a transcrição dos áudios envolvendo a comunicação e também futuras auditorias. Um crescente número de ferramentais vem sendo propostas ultimamente no sentido de se automatizar os processos de reconhecimento de fala, porém estas ainda apresentam limitações de análises, não permitindo uma redundância no processo de transcrição, por exemplo. Propõe-se neste trabalho uma metodologia de análise de áudio em canais separados a partir de gravações de ligações entre os operadores dos sistemas elétricos, visando-se um aumento no grau de robustez da aplicação de processos de reconhecimento fala-para-texto.

Keywords: Speech-To-Text; Speech Recognition; Speech Processing; Whisper Model; Transmission Systems.

Palavras-chaves: Fala-Para-Texto; Reconhecimento de Fala; Processamento de Fala; Modelo Whisper; Sistemas de Transmissão.

1. INTRODUÇÃO

Sistemas de reconhecimento automático de fala (ASR, do inglês *Automatic Speech Recognition*) se baseiam no processo de captação de um sinal de áudio, proveniente da fala do interlocutor, na identificação e reconhecimento das palavras enunciadas e na extração de informações para o uso posterior para a realização de alguma ação ou atividade (Tostes *et al*, 2021), (Barbosa *et al*, 2015). Este tipo de sistema é popular em diversas aplicações, uma vez que a fala é a forma de comunicação mais natural e eficiente realizada entre os seres humanos. Geralmente, o processo de identificação e reconhecimento de palavras realizado pelo sistema se baseia na tecnologia de transcrição de fala para texto (*speech-to-text*), sendo este um outro tema popular de estudos (Malik *et al*, 2021) e usualmente empregado com alguma técnica de aprendizado de máquinas ou inteligência computacional (Nedjah *et al*, 2019).

Devido à sua importância, diversos trabalhos foram desenvolvidos propondo a aplicação de sistemas de ASR nas mais diversas áreas de conhecimento. Todavia, nota-se uma carência de artigos que utilizam a tecnologia de reconhecimento de voz em aplicações de sistemas elétricos

de potência (SEP) (Li *et al*, 2019). Embora existam diversos problemas que podem ser abordados utilizando sistemas de ASR no contexto de SEPs, poucos estudos apresentam destaque na área, sendo a maioria desenvolvida por grupos de pesquisa internacionais, tais como (Li *et al*, 2019), (Zhang *et al*, 2021) e (Zhang, *et al*, 2019). Assim, verifica-se uma lacuna para aplicações que utilizem o processamento de fala do Português, falado no Brasil, para texto.

No Brasil a permissão para o uso de teleassistência no contexto do sistema interligado nacional foi permitida a partir de 2019, pela resolução normativa nº 864 publicada pela Agência Nacional de Energia Elétrica (ANEEL). Todavia, tal normativa foi revogada e substituída pela resolução Nº 1.005, de 15 de fevereiro de 2022 (ANEEL, 2022). No contexto de SEPs, as atividades que devem ser realizadas pelos operadores nos centros de operação são normalmente repassadas para eles por meio de comandos de voz via chamadas telefônicas, como a comunicação dos comandos de operação e despacho de energia que ocorre entre o operador e o Operador Nacional do Sistema Elétrico (ONS). Desta forma, os operadores precisam lidar com altas cargas de trabalho em forma de instruções faladas. Este processo está sujeito a diversas falhas, tais como erros de pronúncia e linhas telefônicas ocupadas (Jorge *et al*, 2010).

Além disso, as operações do dia a dia de um SEP são difíceis e exaustivas, sendo que podem afetar negativamente o desempenho dos funcionários, o que pode então ocasionar falhas, levando-se assim a problemas de segurança (Xiang *et al*, 2021). Quando há a ocorrência de alguma falha, o número de instruções recebidas via comandos de voz aumenta exponencialmente, o que eleva a possibilidade de erros humanos, colocando-se também em risco a segurança do sistema (Yu *et al*, 2020). Ainda, em situações de contingência, é necessário realizar um processo de auditoria das informações repassadas nas chamadas telefônicas, sendo que os engenheiros responsáveis devem ouvir uma grande quantidade de horas de gravação para determinar e identificar o que aconteceu e ocasionou a falha. Durante este processo, muito material irrelevante, no formato de arquivos de áudio, é analisado (Zhang *et al*, 2021).

Por este motivo, sistemas de ASR poderiam ser utilizados para realizar a análise inteligente das informações contidas nos arquivos de áudio, a fim de se determinar quais são relevantes de serem monitoradas para indicar sobre o funcionamento da rede (Xiang *et al*, 2021). Como esta área de reconhecimento ainda é iniciante envolvendo a língua Portuguesa, este artigo propõe uma metodologia de análise a ser utilizada em processos de ASR, visando-se investigar os sinais de áudio destas chamadas telefônicas gravadas em canais de áudio separados, tendo por objetivo principal alimentar as ferramentas de extração de características e com isso melhorar os processos de reconhecimento de fala-para-texto no contexto dos SEPs.

2. ANÁLISE DE ÁUDIOS DE GRAVAÇÕES TELEFÔNICAS

As informações presentes nas chamadas telefônicas realizadas nos centros de operação formam um banco de dados histórico sobre a segurança da rede contendo as falhas e problemas identificados na operação dessa. O uso de tecnologias como as de transcrição de fala para texto permite analisar sistematicamente este tipo de dado, padronizando, rotulando e classificando as informações para uso posterior (Yu *et al*, 2020). Este banco de dados de gravações pode ser apresentado com as falas dos interlocutores gravadas em diferentes canais. O reconhecimento, portanto, das falas para texto com diferentes canais analisados não é usual, uma vez que o mais comum são ferramentas analisarem um único áudio. No caso, as diferentes características vocais dos interlocutores poderiam então gerar imprecisões nas transcrições. Por isso, a hipótese desta pesquisa é verificar a potencialidade da análise de canais separados de gravação para processos ASR.

Os processos aqui apresentados utilizaram o Modelo Whisper de Reconhecimento de Fala-Para-Texto. Este modelo é promissor e pode ser alimentado com palavras em português, porém, ainda apresenta falhas de reconhecimento que podem ser observadas ao comparar as transcrições realizadas pelo Modelo Whisper com as de um especialista humano, como pode ser visto nos exemplos das Fig. 1, 2, 3 e 4, que apresentam o sinal de fala do canal 1 e 2 no domínio do tempo e suas respectivas transcrições.

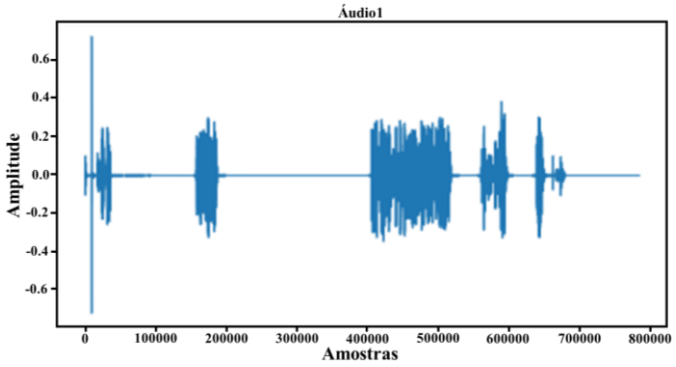

Fig. 1 Amostra de áudio a ser transcrita – Canal 1.

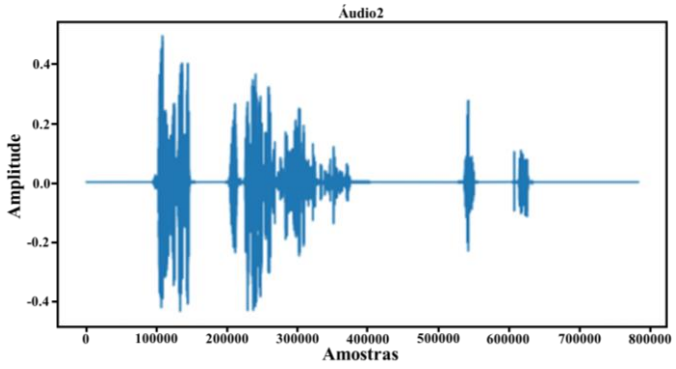

Fig. 2 Amostra de áudio a ser transcrita – Canal 2.

Transcrição especialista humano
"Dimas? Oi Wagner, tudo bem e você? Tá, então na subestação Ariquemes, Ariquemes levar um KV na tensão de referência do compensador síncrono? Atendendo sua solicitação, tá bom Wagner? Falou!"
Transcrição Whisper
"Dimas? Oi Wagner, tudo bem você? Tá, então na substação Ariquemis ele levará um KV na tensão de referência do computador síncrono? Atendendo a sua solicitação, tá bom Wagner? Olá!"

Fig. 3 Áudio transcrito por especialista humano/*Whisper* – Canal 1.

Transcrição especialista humano
"É Wagner ONS centro-oeste, tudo bom? Joia! Seguinte lá em Ariquemes agora as 9 e 57 vamos elevar um KV na referência de tensão do compensador síncrono uno. Positivo. Valeu, obrigado!"
Transcrição Whisper
"Positivo. Valeu, Brilho. Positivo."

Fig. 4 Áudio transcrito por especialista humano/*Whisper* – Canal 2.

Como o Modelo *Whisper* opera de acordo com o contexto da conversa, o seu desempenho de transcrição também foi avaliado a partir de uma segunda abordagem, utilizando-se o áudio original com os dois canais combinados, onde o sinal de fala no domínio do tempo e suas transcrições podem ser vistas nas Fig. 5 e 6, respectivamente.

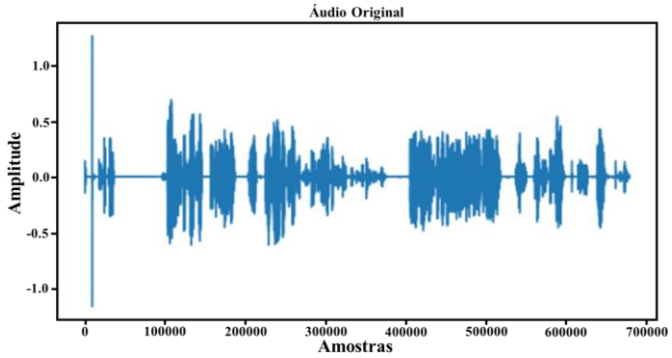

Fig. 5 Amostra de áudio a ser transcrita – Áudio Original.

Transcrição especialista humano
"Dimas? Wagner, ONS centro-oeste tudo bom? Oi, Wagner, tudo bem e você? Joia! Seguinte, lá em Ariquemes agora as 9 e 57 vamos elevar 1 KV na referência de tensão do compensador síncrono uno. Tá, então na subestação Ariquemes, elevar 1 KV na tensão de referência do compensador síncrono? Positivo! Atendendo sua solicitação, tá bom Wagner? Valeu, obrigado! Falou!"
Transcrição Whisper
"Dimas? Wagner, ONS, tudo bom? Oi Wagner, tudo bem você? Jóia! Seguinte, lá em Ariquemis, agora as 9h57, vamos elevar um KV na referência de tensão do computador síncrono 1? Tá, então na substação Ariquemis, elevar um KV na referência do computador síncrono? Positivo! Atendo a sua solicitação, tá bom Wagner? Valeu, brilho! Valeu!"

Fig. 6 Áudio transcrito por especialista humano/*Whisper* – Áudio Original.

3. MÓDULO DE ANÁLISE DE QUALIDADE DE CHAMADA

Com base nos resultados apresentados nas figuras anteriores é possível observar que muito ainda precisa ser melhorado para aumentar a acurácia da transcrição.

Desta maneira, é proposto um módulo de análise para identificar a qualidade da chamada telefônica em relação às interrupções entre interlocutores. Este módulo tem como saídas o índice referente ao grau de interrupção de fala presente na chamada e a indicação dos momentos de interrupção e do interlocutor que causou a interrupção. Na Fig. 7 tem-se uma representação deste módulo proposto.

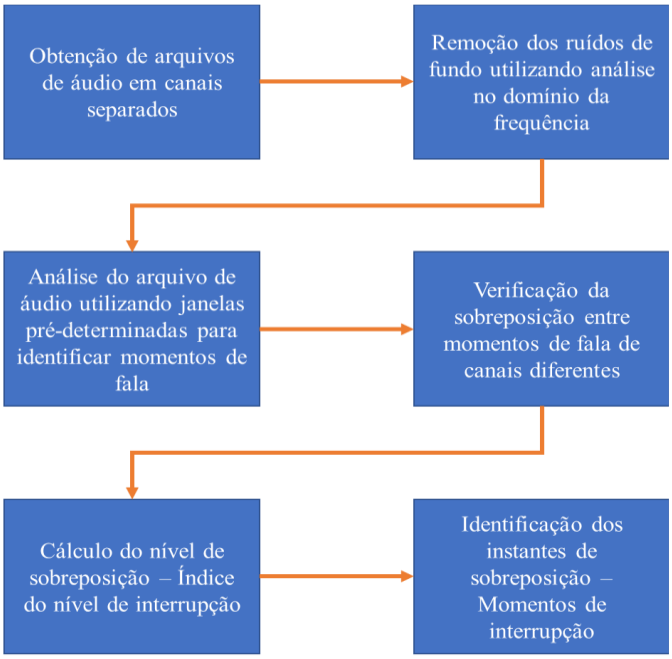

Fig. 7 Fluxograma do módulo de qualidade de chamada.

Na etapa de remoção de ruído é realizada a FFT dos sinais, seguida de uma separação em frames, que são agrupados de acordo com uma comparação dos valores médios no domínio da frequência. Com isso, cria-se uma máscara para separar as frequências vocais com as de ruído. O resultado gráfico deste processamento pode ser observado nas Fig. 8 e 9.

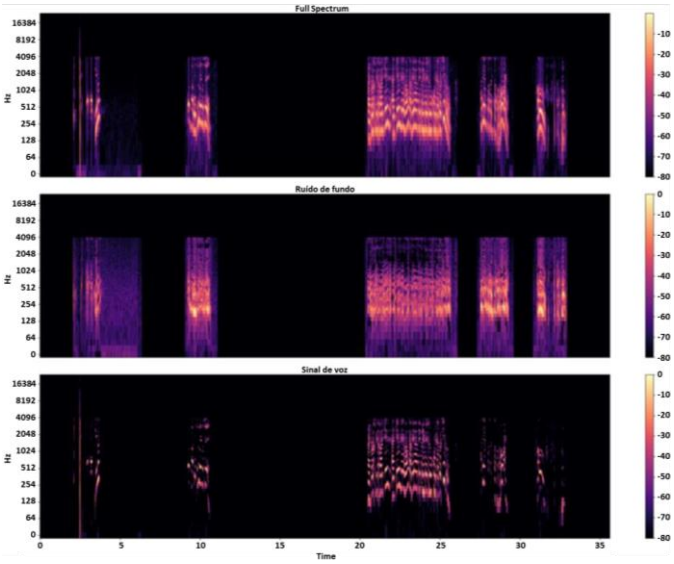

Fig. 8 Análise envolvendo FFT – Canal 1.

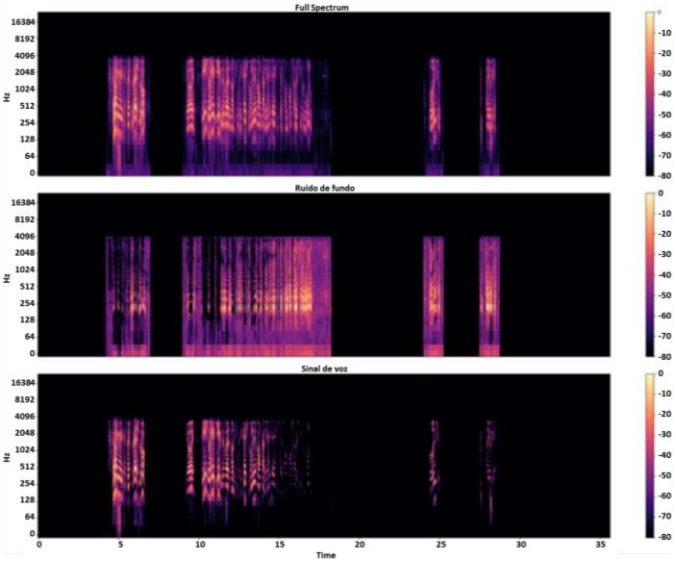

Fig. 9 Análise envolvendo FFT – Canal 2.

Para a detecção de silêncio, uma janela de tamanho determinado em relação à taxa de amostragem do áudio é utilizada para a análise dos momentos de fala e silêncio. O valor RMS do sinal dentro dessa janela é calculado e comparado a um limiar pré-definido, resultando-se em silêncio ou fala. O resultado gráfico desta análise pode ser conferido nas Fig. 10 e 11, onde os retângulos azuis sobrepostos aos sinais representam as posições de fala.

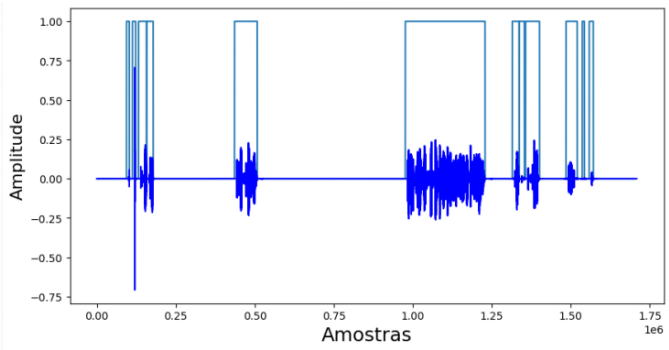

Fig. 10 Análise envolvendo a detecção de silêncio – Canal 1.

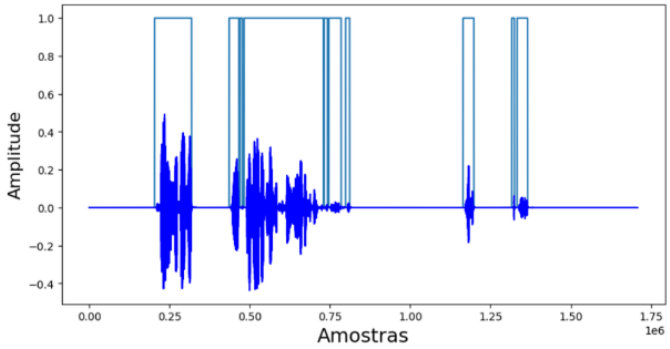

Fig. 11 Análise envolvendo a detecção de silêncio – Canal 2.

4. TESTES DE VALIDAÇÃO

Na Fig. 12 tem-se o exemplo dos áudios das gravações em canais separados.

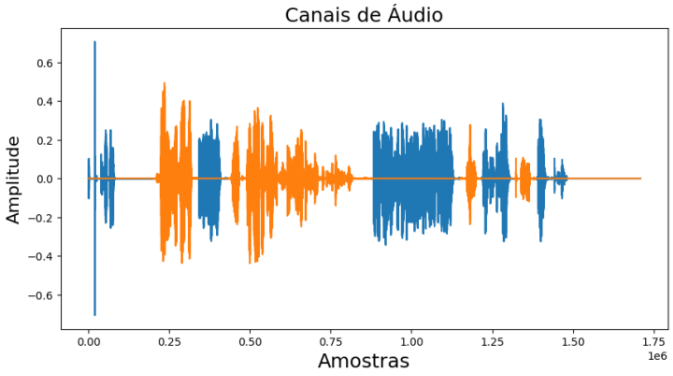

Fig. 12 Áudios em canais separados.

Para a validação do módulo proposto, os sinais foram artificialmente sobrepostos, resultando na informação representada pela Fig. 13.

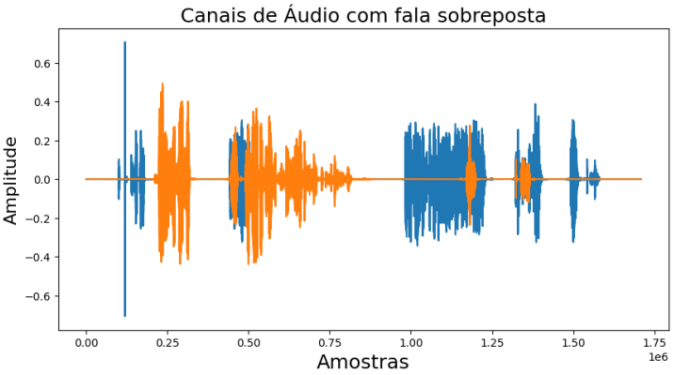

Fig. 13 Áudios sobrepostos.

A identificação dos momentos de interrupção se dá por meio da sobreposição dos intervalos produzidos pelas análises de silêncio, tal como representado pela Fig. 14.

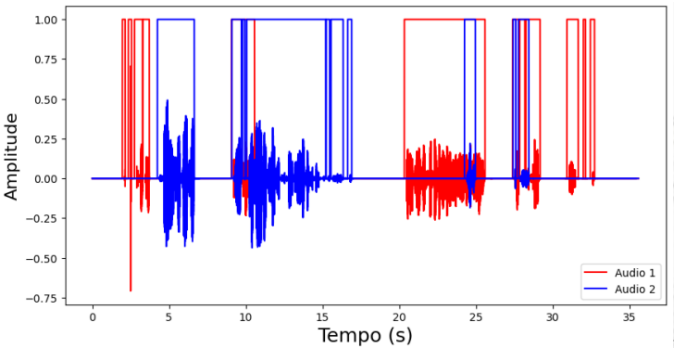

Fig. 14 Áudios identificados com interrupção.

Assim, para a chamada avaliada no desenvolvimento deste trabalho, a taxa de interrupção calculada pelo Módulo de Análise de Qualidade de Chamada foi de 14.33%. Além disso, o módulo ainda fornece como resposta uma análise com todos os momentos de interrupção identificados durante a conversa e indica qual interlocutor foi o responsável por provocar o evento, como pode ser visto na Fig. 15.

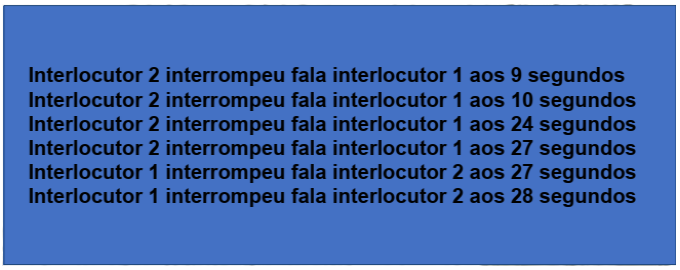

Fig. 15 Momentos de interrupção identificados.

5. CONCLUSÕES

Diante dos resultados de transcrição obtidos, pode-se observar que as duas abordagens apresentam uma boa margem para ter suas transcrições aprimoradas a partir do uso de técnicas de refinamento de aprendizado, fazendo uso de um conjunto de treinamento específico para esta aplicação. Ainda, percebe-se que para a abordagem que utiliza canais separados, o *Whisper* apresenta um desempenho bastante inferior.

Este desempenho inferior é ressaltado na transcrição do áudio do Canal 2, no qual o *Whisper* desconsidera a maior parte do discurso. Isto ocorre, pois, o modelo opera de acordo com o contexto da conversa, por isso, a presença dos dois interlocutores no mesmo arquivo de áudio é essencial para que a previsão da palavra que será escrita possa ser feita dentro do contexto correto.

Entretanto, foi verificado que o uso dos áudios separados em canais distintos nos forneceu informações sobre a qualidade deste processo, permitindo medir o nível de interrupção observado ao longo de uma conversa, a fim de criar uma métrica que indique a qualidade da transcrição com base no sinal de áudio. Além disso, essa abordagem se mostrou vantajosa para identificar momentos de sobreposição de fala, nos quais um interlocutor interrompe o outro, uma vez que nestes momentos o sistema de transcrição apresenta dificuldade em compreender o que está sendo dito.

AGRADECIMENTOS

Este trabalho foi desenvolvido no âmbito do Programa de P&D regulado pela ANEEL, contando também com o apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - CAPES (Processo 88887.636079/2021-00).

REFERÊNCIAS

ANEEL (2022), RESOLUÇÃO NORMATIVA ANEEL Nº 1.005, DE 15 DE FEVEREIRO DE 2022. gov.br, 2022. Available from: <https://www.in.gov.br/web/dou/-/resolucao-normativa-aneel-n-1.005-de-15-de-fevereiro-de-2022-381740251>

BARBOSA, Felipe Gomes; SILVA, Washington Luıs Santos. Reconhecimento de voz para autenticaçao biométrica utilizando máquinas de vetores de suporte e os coeficientes mel-cepstrais. XII Simpósio Brasileiro de Automaçao Inteligente (SBAI). Natal–RN, 2015.

Jiangping, J., Xinjun, Y., Liudong, Z., Xinjie, S., & Zhangliang, S. (2021, December). Analysis of Power Grid Dispatching Instructions Based on BERT-BIGRU Model. In 2021 IEEE Conference on Telecommunications, Optics and Computer Science (TOCS) (pp. 494-499). IEEE.

Jorge, C. A. F., Mól, A. C. A., Pereira, C. M. N., Aghina, M. A. C., & Nomiya, D. V. (2010). Human-system interface based on speech recognition: application to a virtual nuclear power plant control desk. Progress in Nuclear Energy, 52(4), 379-386.

Li, C., Xu, G., Li, W., Lin, W., Ji, T., Zhang, L., & Tang, W. (2020, November). A Novel Speech Recognition Algorithm for Substation Safety Notification Based on IPASS-LSTM. In 2020 International Conference on Smart Grids and Energy Systems (SGES) (pp. 368-373). IEEE.

Li, H., Li, Z., & Rao, Z. (2019, December). Mobile Operation Platform for Power System Maintenance Based on Intelligent Speech Recognition. In 2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS) (pp. 105-108). IEEE.

Malik, Mishaim & Malik, Muhammad & Mehmood, Khawar & Makhdoom, Imran. (2021). Automatic speech recognition: a survey. Multimedia Tools and Applications. 80. 1-47. 10.1007/s11042-020-10073-7.

Nedjah, N., Santos, I., & de Macedo Mourelle, L. (2019). Sentiment analysis using convolutional neural network via word embeddings. Evolutionary Intelligence, 1-25.

Tostes, W. A., Boldt, F. A., Komati, K. S., & Mutz, F. (2021, October). Classificaçao de Sotaques Brasileiros usando Redes Neurais Profundas. In Simpósio Brasileiro de Automação Inteligente-SBAI (Vol. 1, No. 1).

Xiang, Z., Gu, W., Tong, C., Qian, X., Li, Z., & Guo, C. (2021, December). Design of Intelligent Dispatching System Based on Human Voice Adaptive Speech Recognition. In 2021 International Conference on Power System Technology (POWERCON) (pp. 1953-1957). IEEE.

Zhang, H., Xiao, L., Yan, P., & Xiao, Q. (2021, December). Research on Speech Recognition of Power Grid Dispatching Based on Big Data and Deep Learning. In 2021 International Conference on Power System Technology (POWERCON) (pp. 73-78). IEEE.

Zhang, Q., Ma, J., & Wang, J. (2019, November). Application Research of Virtual Operation Robot in Smart Power Plant. In 2019 Chinese Automation Congress (CAC) (pp. 1736-1740). IEEE.

Yu, Y., Guo, Y., Zhang, Z., Li, M., Ji, T., Tang, W., & Wu, Q. (2020, July). Intelligent Classification and Automatic Annotation of Violations based on Neural Network Language Model. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.